

Information Search in Experience-Based Choice and Valuation: Challenges and Applications

**Inauguraldissertation**

zur

Erlangung der Würde eines Doktors der Philosophie

vorgelegt der

Fakultät für Psychologie

der Universität Basel

von

Dirk U. Wulff

aus Frankfurt a. M., Deutschland

Basel, 2015

Genehmigt von der Fakultät für Psychologie

auf Antrag von

Prof. Dr. Ralph Hertwig

Prof. Dr. Jörg Rieskamp

Basel, den 15.1.2015

---

Prof. Dr. Roselind Lieb

## **Declaration**

I, Dirk U. Wulff (born April 18, 1984, in Frankfurt a. M., Germany), hereby declare the following:

- (i) My cumulative dissertation is based on five manuscripts, two of which are published, one of which is currently in revision, and two of which will be submitted shortly. I contributed independently and substantially to all manuscripts. Where I am declared as first author, I was primarily responsible for generating the ideas, collecting the data, and writing the manuscript.
- (ii) I only used the resources indicated.
- (iii) I marked all citations.

Basel, January 30, 2015

Dirk U. Wulff





## **Acknowledgements**

I would like to thank my two advisers, Thomas T. Hills and Ralph Hertwig, for their guidance and support throughout my PhD. I am grateful to my collaborators Max Mergenthaler Canseco and Thorsten Pachur for their input and advice, as well as to Susannah Goss for editing my manuscripts.

I would also like to thank my dear colleagues Renata Suter, Renato Frey and, in particular, Nathaniel D. Phillips, who have made this journey enjoyable, as it was enlightening. Finally, I would like to thank parents, siblings and friends for having carried me this far.



## **Abstract**

How people learn about options may play a more important role than their preferences in determining their choice. This is the bottom line of a 10-year-old research program on the gap between experience- and description-based choice. In this dissertation, I continue this line of research by studying the underlying mechanisms and applying existing knowledge to decision environments in the real world. Paper 1 evaluates the role of recency in the description–experience gap and concludes that it can be traced back to active search strategies, rather than to memory limitations. Paper 2 expands the investigation of recency, showing that its very cause—self-terminated search—poses severe methodological challenges to the study of recency and experience in general. Paper 3 presents evidence for an intricate connection between the length of active information search and preferences, which not only substantiates the claims of the first two papers, but also introduces yet another challenge for the study of experience-based choice. Namely, when a person’s preference determines her information search, the length and outcome of her search may be more indicative of her preference than the choice itself. Taking a much broader perspective, in papers 4 and 5 I apply the knowledge obtained on the description–experience gap to realistic choice environments. Representing a proof of concept, Paper 4 establishes a connection between the description–experience gap and online consumer choices based on different formats of consumer reviews. Paper 5 extends this work by discussing the potential merits of experience-based information formats for private and corporate risky choices and highlights the need to better understand the often intricate relationships between experience and description in real-life choice situations.



## 0 Introduction

Life is full of risk. The choices people make on an everyday basis rarely have certain outcomes. In order to understand human decision making, it is thus imperative to understand how humans deal with risk. For researchers, this means understanding how humans evaluate options by weighing up both the set of possible outcomes and the likelihood with which they occur. Buying a lottery ticket is perhaps the purest example of this. The possible outcomes are that a player may win the jackpot (with a low probability) or win nothing and lose the price of the ticket (with a high probability). Whether a person considers it worth playing will depend on how that person evaluates the options' outcomes in light of their likelihoods.

Numerous studies have investigated how risky choices are made. By analogy to the lottery situation, the vast majority of these studies have asked participants to choose between two or more fully described options with well-defined outcomes and probabilities (Holt & Laurie, 2002; Kahneman & Tversky, 1979); for instance, to choose between:

A: \$4 with probability .8, and otherwise nothing

or

B: \$3 for sure.

Such decisions from description exist not only in casinos, but also in situations in which probabilities are obtained through the aggregation of empirical data, such as the side effects of drugs (Lejarraga, in preparation) or product choices (consumer ratings; Wulff, Hills, & Hertwig, 2014). Often however, tabulated risks are not available—for example, when deciding when to cross the street or where to get lunch. In these cases, decisions may be based on personal experience. The information provided by experience is markedly different to that

offered by explicit descriptions; experience is composed of many individual instances, distributed over time, and inherently limited in number.

Researchers have recently begun to study risky choice in the context of decisions from experience (Hertwig & Erev, 2009; Rakow & Newell, 2010). One approach that has gained particular popularity is the sampling paradigm (Hertwig, Barron, Weber, & Erev, 2004). In this paradigm, no information about the available options is provided upfront. Instead, the options need to be actively explored by participants. To this end, participants repeatedly draw from the available options, each draw being one random sample from that option. As the draws do not entail costs (see Hertwig & Erev, 2009, for other variants), a participant is free to sample from the options however often and in whatever order she likes. When she is ready, she terminates search and indicates the preferred option for one final, consequential choice.

Decisions from description and decisions from experience are alternate ways of learning about one and the same option. Description provides all outcomes with their probabilities (e.g., 4 with probability .8), whereas experience provides a variable sequence of disaggregated outcomes (e.g., 4 ... 4 ... 0 ... 4 ... 0). In principle, an option is the same whether it is presented by description or experience; however, research has repeatedly demonstrated that the two formats lead to systematically different choices (Hertwig et al., 2004). In decisions from description, people choose as if they *overweight* low-probability outcomes, such as the small chance of winning nothing in option A (Kahneman & Tversky, 1979). In decisions from experience, in contrast, people choose as if they *underweight* that same low-probability outcome. For the choice between A and B, these tendencies translated into 36% of respondents choosing option A in the description condition, compared with 88% choosing A in the experience condition. This difference in choice proportions is now commonly referred to as the description–experience gap (Hertwig et al., 2004).

The discovery of the description–experience gap has posed a great challenge to the understanding of how humans deal with risk. Decades of research had informed the construction of intricate and highly successful descriptive models of human decision making, most prominently exemplified by cumulative prospect theory (Kahneman & Tversky, 1997; Tversky & Kahneman, 1992). Yet all of these advances rest on decisions from description. Whether and to what extent these models translate to the context of decisions from experience is a matter of current inquiry. Several mechanisms have been invoked to explain the description–experience gap, but the final verdict with respect to their importance and, to some extent, their existence is in many cases still pending. What seems clear, however, is that the phenomenon of decision from experience is the larger unknown in the study of how humans deal with risk.

One proposed explanation of why description and experience lead to different choices is that limited search alone causes the gap (Fox & Hadar, 2006). Because people have finite resources in terms of time and cognitive capacity, participants in the sampling paradigm must terminate their information search at some point. As samples are random draws from the options, the string of sample experiences resulting from limited search is unlikely to perfectly reflect population-level parameters. Thus, somebody experiencing samples from an option will likely not see one and the same set of outcomes as somebody else studying the description of an option. For instance, a person experiencing option A may by chance see the outcome 0 three times in ten samples, whereas the person relying on description knows that 0 occurs on average in two of ten cases. This sampling error can lead the two to make different choices across the two formats, even if both decision makers have identical preferences. Importantly, contrary to what one might think in the light of random samples, this effect is highly systematic. The binomial distribution that governs the frequencies with which the

outcomes in the risky option A occur is skewed for all events with  $p \neq .5$ . This implies that the majority of sample experiences, as for instance indicated by the median, deviate from the population probability in one direction. For instance, the distribution of occurrences for event 0 in option A is highly right skewed, as a result of which most people will experience its relative frequency of occurrence to be smaller than the actual probability of .2. For this reason, most people will likely choose option A—and will thus choose as if they underweight low-probability events.

The impact of limited information search on decisions from experience is not contested. What is unclear, however, is whether limited information search is sufficient as an explanation to bridge the gap between description and experience. Addressing this question, several studies have investigated the description–experience gap, while ensuring that the information experienced accurately reflects the information presented in the description, either by tweaking the process that generates the samples or by forcing people to sample many times. The results are relatively unequivocal: sampling error diminishes the gap between experience and description substantially, but does not close it (e.g., Ungemach, Chater, & Stewart, 2009). The finding that exposure to equal information in experience and description does not ensure equal choices highlights the need to identify and test other factors contributing to the description–experience gap.

One of the other factors proposed to explain the description–experience gap is recency. When they reported on the discovery of the gap, Hertwig et al. (2004) also observed that choices in the sampling paradigm seemed more influenced by later samples in the sequence than by earlier ones. This greater influence of later samples might not only represent the missing piece in closing the gap; consistent evidence for recency in the sampling paradigm would also imply that experience and description rely on distinct processes. To



date, however, recency in the sampling paradigm is neither well understood, nor has it been consistently demonstrated. Papers 1 and 2 in this dissertation address and substantially advance the issue of recency in the sampling paradigm on the empirical, methodological, and theoretical levels.

Another factor that has been invoked to explain some behavioral patterns in the sampling paradigm—and thus the description–experience gap—is motivation. Hills and Hertwig (2010) observed that the length and pattern of participants’ information search foreshadowed their final choices. This association, so they speculated, may reflect the goals with which participants approach the task. Specifically, collecting only few samples may be an expression of maximizing one’s winnings in the *short run*, whereas collecting many samples may be an expression of maximizing one’s winnings in the *long run*. These two types of choice tendencies, short run and long run, are of course not unique to experience—they are also discussed in the context of decisions from description (Lopes, 1981; Lopes & Oden, 1999). It is thus not the tendency per se, but its link to information search that may contain the key to the description–experience gap. Paper 3 of this dissertation expands on the relationship between choice and search and explains why it represents not only an important aspect in the description–experience gap, but also a major challenge for generally inferring preferences from choices.

The three papers outlined thus far constitute the first part of this dissertation. Beyond their individual focus, they are united by the recurrent theme of active information search. As will become clear, the fact that the sampling paradigm enables participants to actively decide when to terminate information search has profound implications for the interpretation and detection of recency and, for obvious reasons, for the discussion of motivational aspects in experience-based choice.

The second part of this dissertation addresses the relevance of this line of research for decision making in the real world. Despite many open questions, of which this dissertation answers but a few, enough is already known to discuss and test implications of the description–experience gap for real-world situations. Papers 4 and 5 are first attempts in this direction. Paper 4 demonstrates that aggregated and disaggregated information displays of online consumer reviews can influence consumer choice in ways consistent with the description–experience gap for risky choice. Taking a more general perspective, Paper 5 discusses implications of the description–experience gap for corporate risk management.

## **1 The implications and challenges of active information search in experience-based choice and valuation**

In the sampling paradigm of decisions from experience, participants decide when to stop sampling. Surprisingly, research thus far has primarily been interested in the outcomes of such self-terminated information search. Among other questions, studies have assessed the extent of limited information search (Hau, Pleskac, Kiefer, & Hertwig, 2008; Hertwig et al. 2004), which factors influence the length of information search (Frey, Hertwig, & Rieskamp, 2014; Hadar & Fox, 2009; Lejarraga, Hertwig, & Gonzales, 2012; Phillips, Hertwig, Kareev, & Avrahami, 2014; Rakow, Demes, & Newell, 2008), how accurate subsequent frequency judgments are (Camilleri & Newell, 2009; Lejarraga, 2010; Ungemach et al., 2009), and why participants are content to draw relatively few samples (Hertwig, & Pleskac, 2010; Hills & Hertwig, 2010). The psychological and methodological implications of leaving the decision to terminate search to the participant have to date received little attention, however. In paper 1, we show that self-terminated information search may not only explain the behavioral patterns constituting the recency effect, but that it may also be their very source.

**Paper 1: Recency exists in the sampling paradigm of decisions from experience, but why?**

Wulff, D. U., Mergenthaler Canseco, M., & Hertwig, R. (2014). Recency exists in the sampling paradigm of decisions from experience, but why?

Recency in decisions from experience refers to the observation that people more consistently chose the option that later samples indicate to be better than the option that earlier samples indicate to be better (Camilleri & Newell, 2011; Hertwig et al. 2004; Rakow, Demes, & Newell, 2008). However, whether recency implies that later samples truly receive more weight, as has often been assumed (Hertwig et al., 2004; Rakow, Demes & Newell, 2008), or whether it originates from an altogether different process, is yet unclear. In fact, it remains unclear whether recency is indeed a robust phenomenon in the context of choice in the sampling paradigm.

To address these questions, we collected all available data sets obtained using the sampling paradigm and tested them for recency. These data sets included traditional studies, in which sampling was *free*, meaning that the decision to terminate search was left to the participants. However, they also included studies that aimed to minimize deviations between the information experienced and the objectively true option properties, by requiring participants to sample a large, *fixed* number of samples. We found recency to be a robust phenomenon, but only when sampling was *free*. In fixed sampling, neither earlier nor later samples appeared to consistently receive a greater weight.

Recency in the sampling paradigm is commonly thought to originate from limitations in memory capacity, in line with Kareev's *narrow window hypothesis* (Kareev,

2000; Kareev, Lieberman, & Lev, 1997), according to which later samples truly receive more weight because earlier ones are either forgotten or less activated by the time of the decision. If this were the case, however, memory limitations should have played out equally in both *free* and *fixed* sampling. Our analysis—together with other findings in the literature, particularly the finding that frequency judgments elicited from participants after sampling are generally accurate (Camilleri & Newell, 2009a; Lejarraga, 2010; Ungemach et al., 2009) and the lack of association between working memory and any behavioral indicators in the sampling paradigm (Wulff et al., 2014)—suggests that the role of memory in the sampling paradigm needs revision.

Another common interpretation of recency has the potential to explain recency in self-terminated search, namely step-by-step belief updating. In their seminal paper, Hogarth and Einhorn (1992) proposed that there are at least two different modes in which sequences of observations can be processed: *step-by-step updating*, in which new observations are used to repeatedly update a current belief, and *end-of-sequence processing*, in which all samples are first registered and later integrated when a valuation needs to be made (see also Ashby & Rakow, 2014). Hogarth and Einhorn (1992) further showed that the application of these two modes is predominantly a function of the task. When participants were required to produce recurrent evaluations, they exclusively showed recency effects (and no primacy effects); however, when they were required to provide but a single evaluation at the end of the sequence, they mostly showed primacy effects—that is, they gave greater weight to earlier observations. Returning to the sampling paradigm, we suggest that the requirement to self-terminate search will also moderate the use of these two modes of processing. *Free* sampling, which requires the participant to decide when to terminate search, likely causes participants to evaluate the available options during search. *Fixed* sampling, on the other

hand, permits the participant to first observe the sequence of samples before processing them. Thus, *free* sampling may result in recency because the requirement to terminate search triggers a *step-by-step updating* mode, whereas *fixed* sampling does not lead to either recency or primacy effects, because participants are free to choose between the two modes.

Convincing evidence for step-by-step updating would have important implications for the debate on the description–experience gap. Those holding that experience and description are set apart only by sampling error would claim that the sequence of samples is processed into a format akin to description. For instance, the sequence, 4 ... 0 ... 4 ... 4 ... 0, with each outcome weighted equally, would be internally transformed into 4 with a probability of .6, and otherwise nothing, before making a decision. Because step-by-step updating immediately integrates each observation with the current belief and discards the raw sample, such processing must be considered inconsistent with the robust finding of recency in the sampling paradigm.

Yet, as we also show in this paper, a third, previously unconsidered, explanation of recency is consistent with both the finding of selective recency in self-terminated search and the assumption of description-like processing of the raw experience information. It has been argued that participants in the sampling paradigm rely on small samples, because small samples maximize the expected absolute difference between options. Implicit in this argument is that participants act to maximize the difference between the available options. What if they try to achieve this directly by terminating search whenever the difference between the options is large? In this paper, we demonstrate that, as much as belief updating, this behavior can account for recency and its dependence on self-terminated search. Moreover, we highlight that optional stopping does not require any assumptions about how the samples are processed. We argue that as long as the same mechanism is recruited for

search termination and the final choice, recency occurs. On a theoretical level, this proposal reconciles the selective recency effect with the perspective of description-like processing of experience. At the same time, however, it introduces a new layer in the difference between experience and description. If participants are more likely to terminate search when the options appear more distinct, then beyond the effect of limited samples, experiences will systematically deviate from the true options in the direction of larger observed differences. Crucially, our findings in paper 3 indicate that this distortion would not affect differences in terms of the sample mean, but in terms of utilities. For this reason, it is highly difficult to provide direct evidence for this new layer of the description–experience gap.

Before I revisit the issue of preference-dependent, optional stopping in paper 3, I continue to address the question of *step-by-step* versus *end-of-sequence processing* and their detection in a task very similar to the sampling paradigm.

## **Paper 2: Modeling valuations from experience**

Wulff, D. U., & Pachur, T. (2014). Modeling valuations from experience.

What are the cognitive mechanisms underlying valuations based on sequentially experienced samples of a single option's possible outcomes? Ashby and Rakow (2014) have proposed a *sliding window model* (SWIM), according to which people's valuations of an option represent the average of a limited sample of recent experiences (the size of which is estimated by the model) formed after sampling has been terminated (i.e., an *end-of-sequence process*). From their results, Ashby and Rakow conclude that the SWIM performs well relative to alternative models based on model selection criteria. These alternative models included the *value-updating model* (VUM; Hau et al., 2008; Hertwig, Barron, Weber, & Erev,

2006), which is a direct implementation of *step-by-step updating*, and the *summary model*, which assumes no order effects. Further, they report that the individual window sizes estimated by the SWIM correlate with a measure of working memory capacity.

Reevaluating their findings, we highlight several problematic issues in the conclusions drawn by Ashby and Rakow (2014). In a reanalysis of their data, we found no clear evidence in support of any of the models tested, but the *summary model* to show a slight advantage. Further, we demonstrated that individual differences in the window size estimated by the SWIM can reflect differences in judgment noise, rather than true underlying individual differences. In contrast to Ashby and Rakow's evaluation of the data, we conclude that none of their empirical findings support the SWIM. Inspired by Hogarth and Einhorn's (1992) work on order effects in belief updating, moreover, we argue that *end-of-sequence processing*, which the SWIM supposedly reflects, is inconsistent with the active termination of search in this task.

Ashby and Rakow (2014) obtained data from 97 participants, each of whom provided valuations for 40 lotteries, based on samples they had actively sampled. Puzzled by the fact that this rather extensive setup delivered surprisingly little evidence as to which process governs the production of valuations from experience, we further assessed whether study design and methods were at all capable of producing evidence for one process or the other. In order to distinguish the ability of models to account for data, researchers need to work with data that leads to different predictions for the set of models under consideration. Further, when assessing models based solely on their ability to fit existing data, researchers need to know a model's ability to fit data produced by a competing process. As we demonstrate in this paper, both these conditions are severely affected by active, self-terminated information search. First, as in the sampling paradigm, numerous participants sampled very few times,

leading to many strings of observations for which all models necessarily made identical predictions. Second, for reasons within the set of models, some very large sample sizes also led to many strings of observations that resulted in identical predictions. Each of the three models produces, on average, the expected value of a lottery as the valuation, which—due to the law of large numbers—causes the predictions to converge with increasing sample size. Third, the two more complex models, the VUM and the SWIM, operate with different degrees of precision. The VUM has a continuous parameter to account for order effects, which allows the model to fine-tune to observed data, especially when samples are very small. The SWIM, on the other hand, is discrete in that its window size is bound to take one of the discrete steps between a size of 1 and the maximum number of samples collected. The ability to fit any data, also called a model's complexity (Grünwald, Myung, & Pitt, 2005), thus clearly grows with increasing samples size for the SWIM, but may actually decrease for the VUM. Thus, the relative complexity of the two models varies as a function of how many samples were collected.

As we demonstrate in a model recovery analysis, the sum of these issues renders it nearly impossible to identify the true underlying process, at least when standard fit indices are used (AIC and BIC; Burnham & Anderson, 2002), as was done by Ashby and Rakow (2014). For model comparisons where the amount of data (and the information value of that data) is subject to active sampling, the use of more advanced model comparison techniques (Grünwald et al., 2005) and richer experimental designs is advisable. We conclude by making a number of suggestions for methodological tools and features of the experimental design.

Clearly, active, self-terminated search is an important aspect of a task in terms of both the psychological processes involved and their detection. In paper 3, I next return to the issue of preference-dependent search termination.



**Paper 3: How Short- and Long-Run Aspirations Impact Search and Choice  
in Decisions from Experience**

Wulff, D. U, Hills, T. T., & Hertwig, R. (2014). How Short- and Long-Run Aspirations  
Impact Search and Choice in Decisions from Experience. Currently under revision at  
*Cognition*.

To what extent do people adapt their information search and subsequent decisions to their goals? To address this question, we investigated exploration and exploitation policies in choice environments that involved single or multiple plays. In single-play environments, where a choice leads to one random draw from the chosen lottery, the decision maker is forced to make a trade-off between the risk and the potential magnitude of a win. In multi-play environments, on the other hand, the aggregation of risk over multiple draws largely suspends this trade-off. Thus, whereas players in the single-play environment one may feel compelled to choose a small, but safe option, the multi-play environment allows them to safely choose a risky option with a higher long-run expected value. Placing a participant in one environment or the other thus has the effect of switching on and off a person's tendency to avoid risks.

To test this prediction, we conducted an experiment where made several choices between risky options framed as either single or multi-play environments. Results showed that people searched more in the multi-play environment than in the single-play environment. Moreover, the experiences made across these environments differed systematically. Finally, the substantial search effort in the multi-play environment was conducive to choices

consistent with expected value maximization, whereas the lesser search effort in the single-play environment was consistent with the goal of minimizing the risk of winning nothing.

Together, these findings lend evidence to the speculation that people are more likely to terminate their search when, according to their goals (or preference structure), the options appear more distinct. This first account of preference-dependent stopping in the context of the sampling paradigm (see Berger & Berry, 1988, for a discussion of optional stopping in data collection) opens up an entire host of problematic implications. First, and as highlighted before, the long-run expectation in samples from experience will, in many cases, no longer match the expected value of the lotteries. Second, depending on the exact mechanism used to terminate search, the relative location of particular outcomes within the sequence may no longer be random. For instance, a particular undesirable event may be more likely occur toward the end of the sequence, because participants may tend to terminate search after observing it. Third, and probably most importantly, it is no longer clear how to infer the preference of a person in the sampling paradigm.

Let me elaborate on the last point. According to the revealed preference approach (Samuelson, 1938), an individual's preference can be inferred from her choices. Decision problems employing stated probabilities can easily be tailored to make different risk preferences discernible (e.g., Holt & Laury, 2002). In decisions from experience, though, inferring risk preference from choice needs to account for the random composition of samples; it is often the case that no two individuals face identical decision problems. For this reason, it has been argued that an individual's risk preference can only be inferred on the basis of the individually experienced choice environment—and not the objective choice parameters (outcomes and probabilities) (Fox & Hadar, 2006). However, the finding that the environment itself is a function of preference casts doubt on whether even the subjective

environment can be used. Ongoing projects are further investigating the interplay of preference, search, and environment. Although yet unconfirmed, it must be considered a possibility that, in some cases, a person's true preference cannot be reconstructed from her choice, but from the length and pattern of search alone.

### *Interim discussion*

Active sampling is clearly the key to understanding the description–experience gap in the sampling paradigm. Contrary to the focus of previous investigations, however, it is not only the *outcome* of active, self-terminated search that requires attention. I was able to show that leaving search termination to participants has important implications for the psychological models generated to explain choice in the sampling paradigm as well as for the methods used to test those models.

## **2 The description–experience gap in the real world**

Except for visits to the casino, choice options in real life are rarely stated exactly as they are presented in laboratory situations of decisions from description. Likewise, no real-world situation precisely matches of the sampling paradigm in every respect. Nonetheless, choice environments in the real world do resemble the description format, the experience format, and often even both, in important ways. For thousands of years of human history, all that was available to inform a decision was one's own limited experience and the limited experience of others who may be willing to share. Nowadays, with private and public organizations collecting and analyzing massive amounts of data, reliable long-run probability information is becoming increasingly available. Yet in many cases, this information is not easily accessible. And whether decision makers want it or not, past experiences remotely or

directly related to the choice at hand are often available to them. In the light of big data and our knowledge of the description–experience gap, it is thus becoming increasingly relevant to ask how people do and should navigate this multitude of information formats.

Towards these questions, paper 4 attempts to establish a direct link between the two laboratory formats of risky choice and the world of online consumer choice.

#### **Paper 4: Online Product Reviews and the Description–Experience Gap**

Wulff, D. U., & Hills, T. T., & Hertwig, R. (2014). Online Product Reviews and the Description–Experience Gap. *Journal of Behavioral Decision Making*.

Not long ago, buying a book, a music player, or a pair of shoes required consumers to visit their trusted local brick-and-mortar store. Nowadays, many such products are bought online. One driving force behind this development was the invention of online recommendation systems. The initial skepticism associated with anonymous shop personnel and lack of physical contact with the product is now countered by the hundreds of reviews provided by fellow customers. In order to manage the huge numbers of customer reviews, online retailers have developed various formats to summarize the information provided. Amazon.com, for instance, presents customers with a bar plot showing the number of 1- to 5-star ratings a product has been given. In addition, they provide a list of individual star ratings. These two formats, the summary bar plot and the individual ratings, clearly parallel the distinction between description and experience. If consumer choice based on consumer ratings were in any way comparable to risky choice, the two formats could thus have a similar impact on choice.

Assuming that consumer choice based on consumer ratings and risky choice are identical in essential aspects, we tested whether the two formats not only lead to different choices, but do so for the same reasons proposed in previous studies on the sampling paradigm. To this end, we conducted a study in which participants made hypothetical consumer choices based either on an aggregate bar plot of consumer ratings or on a sequence of actively sampled ratings. The results showed that choices differed considerably between the formats. Importantly, they differed due to small sample sizes and recency. These findings establish a promising link between the fields of risky choice and online consumer choice. Research on online consumer choice, although flourishing, is selectively concerned with correlative observations regarding sales figures. The perspective of risky choice can enable the field to go beyond such analyses to study the processes relevant for actual consumer choice. Both the literature on risky choice in general and research on the description–experience gap in particular offer numerous starting points.

Online consumer choice is but one example of real-life situations in which multiple information formats are available. Paper 5 attempts to generalize the implications of the description–experience gap to decision making in the corporate and financial domain.

### **Paper 5: Risky Choice: The Gap between Experience and Description**

Hertwig, R., & Wulff, D. U. (2014). Risikoentscheidungen: Die Kluft zwischen Erfahrung und Beschreibung. *Controlling – Zeitschrift für erfolgsorientierte Unternehmensführung*.

In private and business matters alike, one often can choose how to acquire information about the options at hand. Naturally, the preferred format is always the one that is most beneficial for the decision maker—the format that maximizes the decision maker’s goals. But

which format is that? It is tempting to assume that formats offering reliable probability information (i.e., descriptions from description) are superior to those offering limited sequential information (i.e., decisions from experience). In the light of the consistent overweighting of low-probability outcomes in decisions from description (Kahneman & Tversky, 1979), however, this assumption must be questioned.

In this paper, we discuss consequences arising for private and corporate financial decision making from recent findings on the description–experience gap. In particular, we highlight that the sequential experience of nonlimited information may help to counteract tendencies that often prevent people from making the most beneficial choice. It has been found that, relative to judgments based on description, experience may actually lead to less biased beliefs about the probability of certain events—of course, only if the sequence is a good reflection of the true probabilities (Kaufman, Weber, & Haisley, 2013; Lejarraga, 2010). For both private and corporate decisions makers, this implies that simulation tools affording the sequential observation of many unbiased samples may help to achieve an objective assessment of the choice environment, which can be considered a prerequisite for good decisions.

Ignoring situations in which the information format is under the control of the decision maker, we further argue that both private and corporate decision makers need to always consider the existence of experiences pertinent to a given choice. Experiences may entail accurate judgments of probabilities, but they can still be limited and hence misleading. Little is yet known about the interaction of description and experience, but the few existing studies suggest that the influence of experience is difficult to eliminate, even in the face of clear descriptive warnings in the opposite direction (Barron, Leider, & Stack, 2008).

### **3 General discussion**

The modern information society collects and aggregates information at a higher rate than ever, yet researchers have only begun to understand how people's choices are affected by described aggregate information formats as opposed to individual sequential experience. Limitations in human memory capacity were long believed to separate choice in one format from the other. Yet the findings presented in this dissertation indicate that the impact of memory in the description–experience gap appears to be smaller than has previously been suggested. Far less research attention has been paid to the impact of self-terminated search. I hope I have made a convincing case that the psychological processes involved in experience, the outcome of limited search, and finally the ability to detect certain psychological processes are all strongly impacted by active search termination.

More research is required to fully understand the intricate relationships between search, the experienced environment, and choice. However, if it holds that a person's preference is expressed not only in her choice, but also in her search behavior in terms of preference-dependent stopping, then the research community's image of decisions from experience needs to be revised in many respects.

## References

- Ashby, N. J., & Rakow, T. (2014). Forgetting the past: Individual differences in recency in subjective valuations from experience. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 40, 1153–1162.
- Barron, G., Leider, S., & Stack, J. (2008). The effect of safe experience on a warnings' impact: Sex, drugs, and rock-n-roll. *Organizational Behavior and Human Decision Processes*, 106(2), 125–142.
- Berger, J. O., & Berry, D. A. (1988). Statistical analysis and the illusion of objectivity. *American Scientist*, 76, 159–165.
- Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach*. New York: Springer.
- Camilleri, A. R., & Newell, B. R. (2009). The role of representation in experience-based choice. *Judgment and Decision Making*, 4(7), 518–529.
- Camilleri, A. R., & Newell, B. R. (2011). When and why rare events are underweighted: A direct comparison of the sampling, partial feedback, full feedback and description choice paradigms. *Psychonomic Bulletin & Review*, 18(2), 377–384.
- Fox, C. R., & Hadar, L. (2006). “Decisions from experience” = sampling error plus prospect theory: Reconsidering Hertwig, Barron, Weber & Erev (2004). *Judgment and Decision Making Journal*, 1(2), 159–161.
- Frey, R., Hertwig, R., & Rieskamp, J. (2014). Fear shapes information acquisition in decisions from experience. *Cognition*, 132(1), 90-99.
- Grünwald, P. D., Myung, I. J., & Pitt, M. A. (Eds.). (2005). *Advances in minimum description length: Theory and applications*. Cambridge, MA: MIT press.



- Hadar, L., & Fox, C. R. (2009). Information asymmetry in decision from description versus decision from experience. *Judgment and Decision Making*, 4(4), 317-325.
- Hau, R., Pleskac, T. J., Kiefer, J., & Hertwig, R. (2008). The description–experience gap in risky choice: The role of sample size and experienced probabilities. *Journal of Behavioral Decision Making*, 21(5), 493-518.
- Hertwig et al. 2004
- Hertwig, R., Barron, G., Weber, E. U., & Erev, I. (2006). The role of information sampling in risky choice. In K. Fiedler & P. Juslin (Eds.), *Information sampling and adaptive cognition* (pp. 72–91). New York, NY: Cambridge University Press.
- Hertwig, R., & Erev, I. (2009). The description–experience gap in risky choice. *Trends in Cognitive Sciences*, 13(12), 517–523.
- Hertwig, R., & Pleskac, T. J. (2010). Decisions from experience: Why small samples? *Cognition*, 115(2), 225–237.
- Hills, T. T., & Hertwig, R. (2010). Information search in decisions from experience: Do our patterns of sampling foreshadow our decisions? *Psychological Science*, 21(12), 1787–1792.
- Hogarth, R. M., & Einhorn, H. J. (1992). Order effects in belief updating: The belief-adjustment model. *Cognitive Psychology*, 24(1), 1–55.
- Holt, C. A., & Laury, S. K. (2002). Risk aversion and incentive effects. *American economic review*, 92(5), 1644-1655.
- Kahneman, D., & Tversky, A. (1979). Prospect theory: Analysis of decision under risk. *Econometrica*, 47(2), 263–291.
- Kaufmann, C., Weber, M., & Haisley, E. (2013). The role of experience sampling and graphical displays on one's investment risk appetite. *Management Science*, 59(2), 323-340.

- Kareev, Y. (2000). Seven (indeed, plus or minus two) and the detection of correlations. *Psychological Review*, 107(2), 397–402.
- Kareev, Y., Lieberman, I., & Lev, M. (1997). Through a narrow window: Sample size and the perception of correlation. *Journal of Experimental Psychology: General*, 126(3), 278–287.
- Lejarraga, T., Pachur, T., Frey, R., & Hertwig, R. (2014). Decisions from Experience: From Monetary to Medical Gambles. *Submitted manuscript*.
- Lejarraga, T. (2010) When experience is better than description: Time delays and complexity. *Journal of Behavioral Decision Making*, 23(1), 100–116.
- Lejarraga, T., Hertwig, R., & Gonzalez, C. (2012). How choice ecology influences search in decisions from experience. *Cognition*, 124(3), 334–342.
- Lopes, L. L. (1981). Decision making in the short run. *Journal of Experimental Psychology: Human Learning and Memory*, 7(5), 377.
- Lopes, L. L., & Oden, G. C. (1999). The role of aspiration level in risky choice: A comparison of cumulative prospect theory and SP/A theory. *Journal of Mathematical Psychology*, 43(2), 286–313.
- Phillips, N. D., Hertwig, R., Kareev, Y., & Avrahami, J. (2014). Rivals in the dark: How competition influences search in decisions under uncertainty. *Cognition*, 133(1), 104–119.
- Rakow, T., Demes, K. A., & Newell, B. R. (2008). Biased samples not mode of presentation: Re-examining the apparent underweighting of rare events in experience-based choice. *Organizational Behavior and Human Decision Processes*, 106(2), 168–179.

- Rakow, T., & Newell, B. R. (2010). Degrees of uncertainty: An overview and framework for future research on experience-based choice. *Journal of Behavioral Decision Making*, 23(1), 1–14.
- Ungemach, C., Chater, N., & Stewart, N. (2009). Are probabilities overweighted or underweighted when rare outcomes are experienced (rarely)? *Psychological Science*, 20, 473–479.
- Samuelson, P. A. (1938). A note on the pure theory of consumer's behaviour. *Economica*, 5, 61–71.
- Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, 5(4), 297–323.
- Wulff, D. U., Hills, T. T., & Hertwig, R. (2014). Online product reviews and the description–experience gap. *Journal of Behavioral Decision Making*. Advance online publication. doi:10.1002/bdm.1841

RECENCY EXISTS IN THE SAMPLING PARADIGM OF DECISIONS FROM  
EXPERIENCE, BUT WHY?

Dirk U. Wulff

Max Planck Institute for Human Development, Berlin

Universität Basel

Max Mergenthaler Canseco

Freie Universität Berlin

Ralph Hertwig

Max Planck Institute for Human Development, Berlin

Author Note

Dirk U. Wulff, Center for Adaptive Rationality, Max Planck Institute for Human Development, Berlin, Germany; Max Mergenthaler-Canseco, Department of Philosophy, Freie Universität, Berlin, Germany; Ralph Hertwig, Center for Adaptive Rationality, Max Planck Institute for Human Development, Berlin, Germany.

We are grateful to Susannah Goss for editing the manuscript and we thank the German Research Foundation and the Swiss National Science Foundation for grants to the third author (HE 2768/7-2; 100014-130397).

Correspondence concerning this article should be addressed to Dirk U. Wulff, Center for Adaptive Rationality (ARC), Max Planck Institute for Human Development, Lentzeallee 94, 14195 Berlin, Germany. Phone: +49 30 82406 475. E-mail: wulff@mpib-berlin.mpg.de

### **Abstract**

Life often requires choices between options with uncertain outcomes. How we make these choices has been said to depend on how we evaluate risks. However, recent research has demonstrated that whether we learn about the options available to us through the description of probabilities and outcomes or through the experience of a sequence of individual outcomes also has a considerable influence. One central element in the debate on why experience and description lead to different choices is the recency effect: Recent experiences impact our choices more strongly than earlier ones do; this influence applies only in sequential experiences. In this study, we reanalyze existing data sets to examine the robustness of the recency phenomenon across different variants of the sampling paradigm and different testing methods. Contrary to previous beliefs that recency reflects memory limitations, we find that the recency effect is robust only when participants are required to actively terminate their information sampling. We further demonstrate that this finding is consistent not only with step-by-step belief updating, as has been previously proposed, but also with optional stopping. Which of the two provides the more viable explanation is difficult to discern, but—as we discuss—of critical importance for the interpretation of the description–experience gap.

*Keywords:* description–experience gap, decisions from experience, information sampling, risky choice, recency, optional stopping

## 0. Introduction

If you had the choice between one option offering \$4 with probability .8, or otherwise nothing, and another option offering \$3 for sure, how would you choose? Recent research on choices between monetary gambles has demonstrated that, apart from your personal preferences, your choice would depend on the format in which you learned about the options available (Hertwig, Barron, Weber, & Erev, 2004; Hertwig & Erev, 2009). The main finding is this: When the outcomes and probabilities of a given option are explicitly described, as above, people tend to choose as if they *overweight* low-probability outcomes, such as the chance of winning nothing in the first option (Kahneman & Tversky, 1979; Tversky & Kahneman, 1992). When the same option is directly experienced, however—that is, observed in terms of a sequence of individual outcomes—people tend to choose as if they *underweight* low-probability outcomes. Over the past decade, numerous studies have demonstrated this choice disparity, known as the description–experience gap (Hertwig et al. 2004; Hertwig & Erev, 2009; Rakow & Newell, 2010). Yet the debate as to why this gap occurs and whether choices in the two information formats require independent sets of psychological theories is far from settled. This holds particularly for the sampling paradigm of decisions from experience. In the sampling paradigm, participants first explore initially unknown options in a self-directed sampling phase, taking individual draws from the available options for as long as they like. Once the participant has decided to terminate search, one final, consequential choice is made. In this article, we reevaluate existing data gathered using the sampling paradigm, focusing on one particular causal factor in the description–experience gap, namely recency. As we explain, recency plays a key role in the debate on the description–experience gap in the sampling paradigm because it implies psychological processes that are not relevant in the context of decisions from description.

The description–experience gap in the sampling paradigm has been attributed predominantly to limited information search. In the sampling paradigm, a given option (e.g., \$4 with probability .8, or otherwise nothing) is experienced through the sequential observation of a limited number of random draws (e.g., 4, 0, 0, 4, 4, ... , 4). Unlike decisions from description, where long-run probabilities are given, repeated random draws are subject to stochastic variation. Consequently, the average of multiple sampling sequences will reflect the true underlying probabilities, whereas as individual sequences will not. It is this stochastic error that, for the following reason, leads to the systematic underweighting of low-probability events (Fox & Hadar, 2006; Hertwig & Erev, 2009; Rakow & Newell, 2010): How often an event occurs, given a certain probability and sample size, is governed by the binomial probability distribution. For events with probabilities smaller than .5, the binomial distribution is right skewed, resulting in most people experiencing the event fewer times than expected and fewer people experiencing it more often than expected. The majority of decision makers will thus give the event less weight than it deserves according to its true probability, because they experience the low-probability event fewer times than expected.

The effect of limited information search on choice in the sampling paradigm is consistent with the underweighting of low-probability events: the less people search, the more likely they are to underweight rare events. However, it is unclear whether limited information search is sufficient to fully explain the description–experience gap. One prominent position holds that it does (Fox & Hadar, 2006). According to this approach, participants use the observed relative frequencies to internally construct a description-like representation of an option, which is then subject to the same choice processes that operate in decisions from description. Consistent with this position, studies have found that cumulative prospect theory, the benchmark model of decisions from description (Erev et al., 2010; Kahneman & Tversky,

1979), is relatively successful in describing choices in the sampling paradigm (i.e., choices based on the relative frequencies found in subjective experience), but entirely fails to describe choices based on the objectively true probabilities (Fox & Hadar, 2006). What is implicit in this position and specifically the reliance on relative frequencies, however, is the assumption that each of the outcomes experienced receives equal weight, regardless of when it was experienced. In other words, the argument assumes no order effects.

Recency implies that observations made later in the sequence receive more weight than do those made earlier. When some observations are given more weight than others, this has the effect of reducing the effective sample size. For illustration, let us assume that all weight is placed on a subset of samples. In this case, the sample size will effectively be reduced by the number of samples receiving no weight. Thus, recency causes underweighting of low-probability events for the same reasons that limited information search does. Importantly, however, by reducing the effective sample size, the effect of recency is that of the effect of limited information. That is, even if the true probabilities are accurately reflected in the total sequence of samples, this will not necessarily be the case for the subset of samples on which choices are based.

Like underweighting of small probabilities, weighting in the recency phenomenon is not directly measured; it is inferred from the pattern of choices. Specifically, when earlier samples indicate that one option is better but later samples indicate that the other is better, choices consistent with *later* samples are said to reflect recency. The debate as to which psychological processes cause recency—and indeed whether the phenomenon exists at all—is currently ongoing. The empirical findings to date are mixed (see Table 1). Some studies, including the original study by Hertwig et al. (2004), have shown recency. However, an equal



number of studies have found no effects, and a few studies have found choices to be more consistent with earlier samples, implying the opposite of recency, which is called primacy.

Table 1. *Summary of Previous Findings on Order Effects (see Appendix A for Study-by-Study Details)*

	Sampling type ( <i>n</i> data sets)			Total %
	Free (7)	Matched (4)	Fixed (8)	
Recency	5	—	3	<b>42</b>
No effect	2	4	3	<b>47</b>
Primacy	—	—	2	<b>11</b>

*Note.* Free, matched and fixed sampling are variants of the sampling paradigm. Free sampling refers to the version introduced thus far. In matched sampling samples are not randomly generated, but from an underlying algorithm that aims to minimize the difference between the objective problems and the subjective experience. Fixed sampling includes all free and matched sampling procedures, in which the sample size is predetermined by the experimenter. Further explanations follow below in the text.

To clarify the role of recency in the sampling paradigm, we reanalyze the existing data obtained using different variants of the sampling paradigm and different methods to test for recency. If our findings show that recency is a robust phenomenon across these conditions, then this indicates that decisions from experience and decisions from description are set apart by more than mere sampling error. If, in addition, we find evidence for one of the more psychologically grounded explanations of recency, this will call for the development of independent theories for the two information formats. Against this background, we next review the empirical evidence for three different theoretical explanations of the phenomenon of recency in the sampling paradigm: memory limitations, step-by-step belief updating, and optional stopping.

## 1. A review of previous explanations of and evidence for recency

Explanations of recency in the sampling paradigm have focused primarily on the role of memory limitations (Atkinson & Shiffrin, 1968; Murdock, 1962). In particular, Kareev's *narrow window hypothesis* (Kareev, 2000; Kareev, Lieberman, & Lev, 1997) assumes that inferences about multiple pieces of information are based on as many pieces as can be held in working memory at a single point in time. Because working memory is known to be limited (Cowan, 2001; Miller, 1956), the narrow window hypothesis leads to the prediction that participants either constrain their sampling effort to their working memory capacity or consider only the most recently collected information in making their choice.

In support of the limited capacity explanation, Rakow, Demes, and Newell (2008) found that the length of information search was positively related to working memory span, suggesting that participants adjusted their information search to their processing capacities. Further, Rakow and Rahim (2010) found that children, whose working memory capacity is presumably not yet fully developed, showed more pronounced recency effects than did adolescents or adults. However, the evidence *against* a memory-based interpretation of recency in the sampling paradigm weighs heavier. First, Rakow et al. (2008) found that participants with different working memory spans did not differ in the extent to which they showed a recency effect. Second, other studies measuring working memory capacity have been unable to replicate a correspondence of working memory and search behavior in the sampling paradigm (Wulff, Hills & Hertwig, 2014a; Wulff, Hills, & Hertwig, 2014b). Third, studies that probed participants' memory by asking them to estimate the observed frequency of outcomes after sampling found high levels of accuracy (Camilleri & Newell, 2009a; Lejarraga, 2010; Ungemach, Chater, & Stewart, 2009). Last but not least, in variants of the sampling paradigm created to rule out the effect of sampling error, recency was found to be

substantially diminished. Table A1 in Appendix A presents these findings in detail. In *matched* sampling, where underlying algorithms ensure that the observed outcomes match the true probabilities, all studies that tested for recency failed to find an effect. In *fixed* sampling, where sample size is predetermined by the experimenter, findings were mixed, with some indicating recency and others primacy. The lack of recency in matched sampling is likely an inevitable consequence of the method, with no conceptual implications.<sup>1</sup> Importantly, however, memory limitations should have been observed in fixed designs. Although Rakow and colleagues (2008) argued that liberating participants from terminating search may have reduced the necessary effort and, consequently, the constraining effect of memory, it seems more plausible that the role of memory limitation in the sampling paradigm is itself limited.

Another class of explanations focuses on the sequential nature of information processing in decisions from experience. Most prominently, Hertwig et al. (2004) proposed that recency may arise from a process of step-by-step belief updating. According to models of this class, new observations are not stored explicitly in memory, but are directly integrated with the running value of the respective option by means of an anchoring-and-adjustment process (Hogarth & Einhorn, 1992; March, 1996). Specifically, such models first determine the difference between the current belief, which serves as the anchor, and the new observation. They then update the belief by computing a weighted average of this difference and the current belief. As shown by Hogarth and Einhorn (1992; see also Anderson & Hovland, 1957), updating in this way always leads to recency if the new observation receives a weight larger than zero. What is appealing about this alternative explanation for recency in the sampling paradigm is that the mechanism is not assumed to be universally applied.

---

<sup>1</sup> Experiments using matched sampling may well have failed to find an effect because the hidden sampling algorithm ensures that all subsets of samples are equally representative of the underlying true probabilities, which renders the predictions of early and late samples indistinguishable (Camilleri & Newell, 2011b).

Updating in this way is thought to capture behavior only if participants engage in step-by-step processing of the observed information, which may be because they chose to or because they are required to provide intermediate valuations. Precisely this circumstance may reconcile the conflicting findings for free and fixed sampling. When sampling is free, and participants themselves decide when to terminate sampling, they are likely to form valuations of the available options during sampling to choose an appropriate place to stop. In other words, unless participants internally apply a fixed sampling regime, they need to assess whether they have collected enough information, most likely by repeatedly determining some utility and/or uncertainty measure for the available options (see Wulff & Pachur, 2014, for a more extensive argument). This reasoning clearly implies that self-terminated search requires some form of online processing, which could be implemented as a step-by-step belief-updating process. When sampling is fixed, however, and participants do not have to decide when to terminate search, they are free to merely observe the incoming information until they are required to make their choice. According to the belief-updating framework of Hogarth and Einhorn (1992), such end-of-sequence processing gives rise to primacy, rather than recency. This very freedom to choose one strategy or the other may explain why studies in which sample size was fixed by the experimenter found evidence for both recency and primacy.

Both of the explanations for the recency phenomenon presented thus far—memory limitations and belief updating—focus on what information from a sequence of samples is integrated, and how. A third, as yet unexplored explanation, is that the true reason for recency lies not in information processing, but in the process of information search. Investigating the reasons for relatively early search termination in the sampling paradigm, Hertwig and Pleskac (2010) argued that participants might capitalize on the fact that the expected absolute difference between the means of the observed options is larger for small than for large sample

sizes. In other words, people may restrict their search to a few samples because it makes choice easier. By the same token, it seems plausible that participants may also terminate their search when a choice appears easy. As we demonstrate in detail below, such optional stopping tendencies will show up as recency in all existing methods devised to test for recency. Importantly, because optional stopping requires the participants to terminate search themselves, this will play out only in free sampling, but not in fixed sampling.

To summarize, we have outlined three potential explanations of recency: memory limitations, step-by-step belief updating, and optional stopping. The mixed results for fixed sampling methods favor the latter two explanations. However, it remains unclear whether recency is indeed a robust phenomenon in the sampling paradigm and, specifically, whether it is predominantly found for free as opposed to fixed sampling (Hertwig & Erev, 2009). In a first step to close this gap in the research, we next reanalyze all existing data sets using the sampling paradigm.

## **2. A meta-analysis of recency in the sampling paradigm**

### *Data collection*

To evaluate the robustness of recency and test the proposed explanations of recency, we sought to collect all existing data sets using the sampling paradigm. To this end, we first searched through all articles that the *Web of Science* (Thomson Reuter's) and *Google Scholar* (Google Inc.) identified as citing one of the three original articles addressing the impact of experience on risky choice: Barron and Erev (2003), Hertwig et al. (2004), and Weber, Shafir and Blais (2004). We found a total of 1234 articles. Of these, we selected all those that fulfilled the following criteria: (1) *Prior ignorance*: Prior to search, no information about the underlying probabilities was available. (2) *Active search*: Participants engaged in active

information search (i.e., participants' behavior was the cause for the generation of individual draws from the selected options). (3) *Inconsequential search*: Aside from inevitable opportunity costs, the length or the order of search was not associated with either costs or benefits. (4) *Probabilistic options*: At least in the eye of the participant, the probabilities of the available options were distributed across at least two outcomes. Application of these criteria left us with a total of 27 articles. We contacted the authors of these articles to request the original data and obtained usable data for 23 studies. In three cases, the study did not record the sequence of samples drawn by participants. In one case, the data had apparently been lost. To obtain a comparable set of studies, we further reduced the set to studies that required active exploration of two two-outcome lotteries. This led to the exclusion of one study in which the number of outcomes per option ranged from three to five, and to the exclusion of two studies in which the risky option, but not the safe option, had to be actively explored. We further added the data from one unpublished study by the first and third authors (Wulff & Hertwig, 2014) that aimed to replicate the original study by Hertwig et al. (2004) in an online sample acquired through the Amazon Mechanical Turk (Paolacci, Chandler, & Ipeirotis, 2010).

The final set of 21 articles was then split into 53 units of observation for the analyses. Units were either independent studies within an article or meaningfully different conditions within a study. For instance, we separated out the subsamples in Frey, Hertwig, and Rieskamp (2014) in which different moods were induced prior to application of the sampling paradigm. In total, we identified 31 units using free sampling, 4 units using matched sampling, and 18 units using fixed sampling. Table 2 gives an overall summary, and Table A2 in Appendix A provides details of the studies and the units, including the number of participants, the type of lotteries used, and the median sample size drawn by participants.

Table 2. *Summary of Data Used for the Reanalysis of Recency*

	Units	Articles	<i>N</i> Participants	<i>N</i> Trials
Free	31	16	1611	12336
Matched	4	2	226	977
Fixed	18	5	688	2700

*Data analysis*

Four methods have been used to test for recency effects in the studies identified, three of them inferring recency from choice in similar ways. The general principle is this: If a choice is consistent with the better option in later samples, it is said to reflect recency. Analogously, if a choice is consistent with the better option in earlier samples, it is said to reflect primacy. Of course, this raises two questions: (a) What is considered to be the better option? (b) What counts as early and late samples? All three methods identify the better option as the one offering the higher mean in the respective subset of samples. The methods differ, however, in how they assign samples to early and late subsets. As illustrated in Figure 1, the *option-split* method, which is predominant in the literature (see Table 1A), splits the samples option-wise. That is, the first half of the samples for each option is used for the primacy prediction, the second half for the recency prediction. The *problem-split* method splits the samples in half across the entire sequence. That is, if 12 samples were collected, the first 6 are used for the primacy prediction and the later 6 for the recency prediction. The *switch-split* method splits the sequence of samples along transitions (switches) between options. Samples around the first switch (i.e., all samples before the second switch) are used for the primacy prediction; samples around the last switch (i.e., all samples after the second-to-last switch) are used for the recency prediction.

(1) Option-Split

A	3	3	3		3			3	3				
B				0		0	0				0	32	0

(2) Problem-Split

A	3	3	3		3			3	3				
B				0		0	0				0	32	0

(3) Switch-Split

A	3	3	3		3			3	3				
B				0		0	0				0	32	0

Figure 1. Illustration of the methods applied to infer recency in the sampling paradigm.

Primacy predictions are based on beige cells, recency predictions on turquoise cells.

Despite their similarities, the three methods have different advantages and disadvantages. The *option-split* method is robust against variations in sample size; separate predictions can be made for recency and primacy for most sequences. However, it is highly sensitive to the order in which samples are taken. For instance, if a participant samples only once, the samples used to derive the primacy prediction for one option may have been collected after those used to derive the recency prediction for the other option. This issue is solved when samples are split problem-wise in the *problem-split* method. With this method, primacy and recency predictions are always cleanly separate from each other. The downside, however, is that it is only possible to make predictions for either recency or primacy if the participant switched only very few times. Finally, the *switch-split* method combines the virtues of the other two methods; it is computable for all sequences and nicely separates primacy from recency in terms of time. The problem with this method is that the number of samples used to derive primacy and recency predictions may differ drastically depending on



the switching pattern. See Appendix B for an algorithmic description of the three methods, including the handling of overlaps and failures.

In the following, we use each of these methods to evaluate recency effects in the existing data. One reason we use all three methods is to test for robustness. If recency is found using all three methods, then its detection or lack thereof cannot be attributed to the downsides of any of the methods. Another reason is that the differences between the methods may help to illuminate previous findings in the literature. For instance, using the *option-split* method, Rakow and Rahim (2010) found that the degree of primacy was related to how frequently people switched between options during search. This finding might be explained by the fact that, in the *option-split method*, both recency and primacy rely on different subsets of the sequence for frequent and infrequent switchers. This suggests that the inferred order effect may not be found consistently across the methods. Moreover, Camilleri and Newell (2011b) found recency in a fixed design when using the *problem-split* method, but not when using the *option-split* method, suggesting that it is a sample's absolute and not relative position that determines its weight. Finally, two articles have suggested yet another possible reason for the finding of recency, namely that participants perform a two-step exploration of the options, acquiring an overview of the options before actually estimating the probabilities of the outcomes (Camilleri & Newell, 2009a; Rakow et al., 2008). Such a two-step mechanism would presumably best be detected by the *switch-split* method, because the samples around the first switch would likely be used for exploratory purposes only. Thus, if participants do take a two-step approach, more extreme findings should emerge for the *switch-split* method than for the other two.

A fourth method that has been used to test for recency is the value-updating model (VUM; Hau, Pleskac, Kiefer, & Hertwig, 2008; Hertwig, Barron, Weber, & Erev, 2006). It

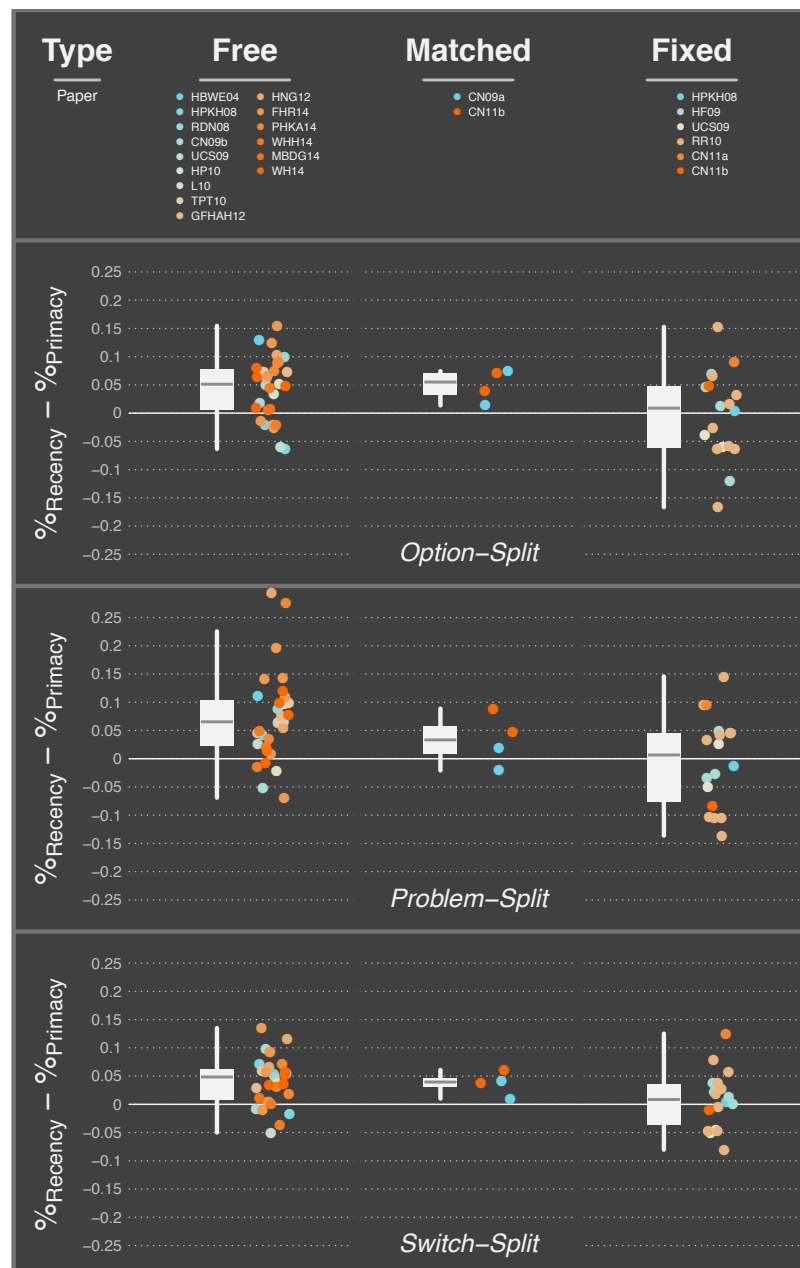
differs from the previous three models in that it represents a more complex measuring model, allowing for more fine-grained differences in recency or primacy. It rests on the same logic as the *option-split* method. Where the *option-split* method weights the sample of one option according to a step-function ranging from 0 to 1, however, the VUM weights outcomes in a gradually increasing or decreasing manner. Moreover, instead of assuming particular weights, as done by the other three methods, the VUM can be fitted to the data. Clearly, the VUM has desirable properties. Because it is based on an option-wise assessment of recency and primacy that is already covered by the *option-split* method, however, we decided not to include it in the analysis. In contrast to results from the VUM, results from the *option-split* method are directly comparable to those obtained by means of the other two methods, predominantly because it also has no free parameters.

Using the three methods—*option-split*, *problem-split*, and *switch-split*—we analyzed, for each individual participant and decision problem, whether the choice was consistent with the higher mean in samples assigned recency and/or in samples assigned primacy. For the statistical analyses, we then aggregated the findings on the level of participants, by computing the difference between the percentage correctly predicted by recency and primacy, respectively. For the display in Figure 2, we computed the same difference, but on the level of the unit.

## Results

Figure 2 shows the results of our recency analysis. The top panel lists the studies that applied the free, matched, and fixed variants of the sampling paradigm, in the left, middle, and right column, respectively. The other three panels show the results for the three methods, in terms of the difference score reflecting the difference in percentage points between the

recency and primacy prediction. Values higher than 0 indicate a recency effect—that is, that recency accounted for more choices than primacy did—and vice versa. Each point represents a unit of observation; its color indicates the respective article. The gray boxplots to the left summarize the results for the respective combination of recency method and sampling variant.



*Figure 2.* Unit-level results of the meta-analysis on recency. The top panel indicates which articles contributed data to the free, matched, and fixed sampling conditions respectively. The

other three panels show the results of the analyses for the three methods applied, separately for the three sampling variants. Points reflect unit-level difference scores that are summarized in the adjacent boxplot.

Is recency a robust phenomenon? The results indicate that it is. The data collected under free sampling show that the majority of data points lie above 0, indicating a recency effect. In sum, the three methods led to a median advantage of 4.8 to 6.5 percentage points for recency over primacy across the three methods. Linear mixed-effects analyses that regressed the difference in percentage points for each individual on a general intercept, while accounting for repeated measurements at the level of the unit and article through random intercepts, indicated significant deviations from zero for free sampling in all three methods (see Table 3). The same pattern, but somewhat reduced in magnitude, was also found for matched sampling: Median advantages for recency amounted to 3.3 to 5.5 percentage points, which corresponded to significant deviations from 0 in the *option-split* and *switch-split* method.

Critically, did the same pattern also hold for fixed sampling? No, the data indicate that fixed sampling led to neither a robust recency effect nor a primacy effect. On aggregate, the data indicate a meager advantage of 0.7 to 0.9 percentage points for recency, which did not correspond to a significant deviation from 0 in any of the methods. It must be noted, however, that fixed sampling led to considerably larger variation than did free or matched sampling. In themselves, individual units deviated clearly from 0, the point of indifference, particularly the units of Rakow and Rahim (2010). This finding may suggest that recency or primacy is possible under fixed sampling if certain conditions are met. On the other hand, the age groups in the three experiments that make up the units of observation in Rakow and Rahim (2010)

are comprised of relatively few participants relative to other units (see Table A2 in Appendix A), which may also account for the high variation in results for fixed sampling.

The analysis showed no clear evidence that the different methods led to different results. For instance, although the *problem-split* method made predictions for notably fewer cases (61% to 69%) than the other two methods did (89% to 97%; see Appendix B for details of strategy implementation), no marked differences between the methods were observed. The lack of an impact of the method used was further confirmed when we computed the results only for cases in which all three methods led to valid predictions. As shown in the last column of Table 3, this approach generally resulted in an increased recency effect, but it did not show that one method was consistently more sensitive than another.

Table 3. *Results of the Meta-Analysis on Recency*

Variant	Method	Median $\%_{\text{Rec}} - \%_{\text{Prim}}$ (acc. Fig. 2)	Mixed effects analysis (against 0)	% valid	Median $\%_{\text{Rec}} - \%_{\text{Prim}}$ (restricted)
Free	Option-split	5.1	$t(11) = 4.56$ $p = .002$	92	8.3
	Problem-split	6.5	$t(24) = 5.40$ $p < .001$	65	7.4
	Switch-split	4.8	$t(13) = 5.44, p < .001$	91	7.4
Matched	Option-split	5.5	$t(648) = 2.65$ $p = .008$	97	6.9
	Problem-split	3.3	$t(188.9) = 1.68$ $p = .095$	61	4.2
	Switch-split	3.9	$t(224) = 2.40,$ $p = .017$	96	5.8
Fixed	Option-split	0.9	$t(648) = .18$ $p = .860$	91	0
	Problem-split	0.7	$t(588.6) = -.15$ $p = .880$	69	-2.6
	Switch-split	0.9	$t(4) = .26,$ $p = .806$	89	1.1

*Interim summary*

Our analysis using three established methods to test for recency showed a robust recency effect for free and matching sampling, but not for fixed sampling. As has been speculated in the literature, recency thus occurs in the sampling paradigm only in self-terminated search. When considering only the two classic interpretations of recency—memory limitations and step-by-step updating—this finding must be seen as evidence of step-by-step updating. If memory limitations were involved in producing recency, consistent signs of recency should also have been found for the fixed sampling variants. As additional analyses revealed, even in those studies that required participants to sample up to 80 to 100 times (specifically, Camilleri & Newell, 2001 1a; Hau et al., 2008; Ungemach et al., 2009), an aggregate advantage of just 0.1 percentage points was found for recency by the *option-split* method. Thus, the argument of reduced effort in fixed sampling, which has been put forth to save limited memory capacity as a viable explanation of recency (Rakow et al., 2008), does not seem to hold either.

Finally, we found no evidence for the two minor ideas that the absolute position of a sample is more important than its relative position, or that initial samples are discarded for the final evaluation, as implied by the proposed two-step exploration process (Camilleri & Newell 2009a; Rakow et al., 2008).

**3. Optional stopping as an alternative interpretation for recency**

Before we can conclude that recency is a result of step-by-step updating, however, we need to address another interpretation outlined in the introduction: optional stopping. What distinguishes free from fixed versions of the sampling paradigm is not only that the free

version can be expected to elicit information processing before search ends, but also that participants can decide when to terminate search. Thus far, the literature on the sampling paradigm has largely neglected to consider when participants terminate search—as reflected by the assumption that the binomial distribution can explain the aggregate pattern of experiences due to limited search (Hertwig et al., 2004; Hertwig & Pleskac, 2010). The binomial distribution only provides the probability distribution of the number of occurrences given a particular sample size. That is, in order to apply the binomial distribution, researchers need to presuppose a certain sample size, which translates into assuming that participants terminate search either at random (as explicitly assumed by Gonzalez & Dutt, 2011) or at a fixed, predetermined point in the sequence. Contrary to this assumption, however, research has found that sample size scales with the variance of lotteries, which implies that participants adapt their sample size as a function of what is observed (Lejarraga, Hertwig, & Gonzalez, 2012). In this case, which experiences lead to the termination of search? Neglecting variance sensitivity for now, one obvious possibility that resonates with the amplification effect proposed by Hertwig and Pleskac (2010) is that participants terminate search when the difference between options seems large. The amplification effect proposes that participants sample only little, because small samples tend to maximize the absolute difference experienced between options. Thus, if participants act to maximize the difference between the options, they may well do so by stopping when the difference seems large.

To see how optional stopping could lead to recency, let us consider a person who has experienced the sequence illustrated in Table 4 and is about to collect one more sample from option A. If she samples one more time, the most likely outcome is 4, which will, given that she was about to terminate search, probably amount in the choice of option A. But what if the other event occurs, and she makes the surprising observation of 0 in the last sample? What

will she do now? One possibility is that she terminates search nonetheless and chooses A, which is still the better choice in terms of sample means. In that case, her choice will be consistent with primacy. However, it seems much more likely that she decides not to terminate search, but to collect more samples. In that case, the outcomes of the additional samples will strongly influence her choice. Importantly, irrespective of whether the next few outcomes speak in favor of A or B, she will likely choose in accordance with them, leading to a choice consistent with recency. Thus, one way or another, recency will prevail if final samples that go against a previously formed opinion lead to more information search.

Table 4. *Exemplary Sampling Sequence for a Choice Between \$4 with  $p = .8$ , and Otherwise Nothing, Versus \$3 for Sure*

Option				Samples										
A				4	4	4						4	4	?
B	3	3	3				3	3	3	3				

Against this background, we conducted a simulation to test the effect of random and optional stopping on the results of the option-split method. Specifically, we generated for both sampling strategies 10,000 agents that each played the sampling paradigm six times, once for each of the problems introduced by Hertwig et al. (2004). To lend the simulation some realism, we implemented random sampling not according to a uniform distribution, but dependent on the number of samples collected so far. Specifically, an agent in the random sampling condition initially drew one sample from each option and then decided sample-by-sample whether to terminate search according to:

$$p(\text{terminate}) = 1 - \left(\frac{1}{n^{.01}}\right), \quad (1)$$

where  $n$  denotes the current sample size. This implementation not only reflected the heavily right-skewed sample size distribution found for empirical data, it also ruled out any



differences between optional and random stopping due to variable stopping points. For agents in the optional stopping condition, we based the decision to terminate search on the exponential choice rule (Rasch, 1980) that also eventually determines the agent's final choice:

$$p(\text{choose } A) = (1 + e^{(\bar{x}_B - \bar{x}_A)})^{-1}, \quad (2)$$

where  $\bar{x}_A$  denotes the sample mean for option A and  $\bar{x}_B$  the sample mean for option B. This rule, which is used by both conditions to terminally decide between the options, converges to a probability of 1 when  $\bar{x}_A$  grows relative to  $\bar{x}_B$  and to 0 when  $\bar{x}_B$  grows relative to  $\bar{x}_A$ . Search termination for optional stopping thus becomes more likely, the larger the differences between the options are. To prevent search being terminated each time the first two samples contain a positive value for one option and a 0 for the other option, we scaled the difference of sample means with  $n/100$ . The impact of the difference between sample means thus grows linearly with sample size, from 1/100 times the difference for a sample size of 1 up to the absolute value of the difference for a sample size of 100—by which time most agents will have terminated search already. The process to terminate search thus first determines the probability that the agent would choose option A given the difference in sample means and the sample size (note that  $p_B = 1 - p_A$ ):

$$p_A = \left(1 + e^{\left[\frac{n(\bar{x}_B - \bar{x}_A)}{100}\right]}\right)^{-1}. \quad (3)$$

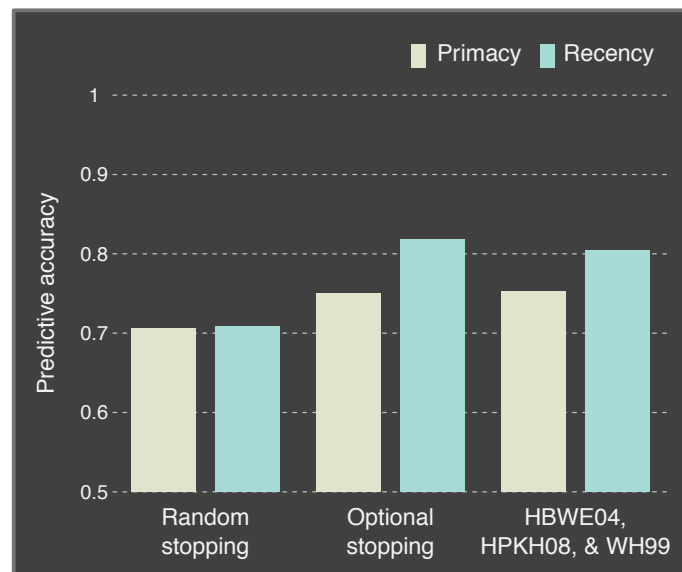
The agent then evaluates the extremity of that choice probability and terminates search according to:

$$p(\text{terminate}) = 2 * (\max(p_A, 1 - p_A) - .5), \quad (4)$$

which is 0 when the choice probabilities for both options are identical and 1 when the choice probability for one of the options is also 1. Note that this implementation of optional stopping is consistent with example presented in Table 4, but not a verbatim reflection; we will return

to this issue in the Discussion. After search has been terminated, agents choose according to the probabilistic choice rule given by (2).

Figure 3 plots the results for random stopping and optional stopping alongside the empirical results of three studies that implemented the six Hertwig et al. (2004) problems in a free sampling paradigm (Hau et al., 2008; Hertwig et al., 2004; Wulff & Hertwig, 2014). Specifically, the figure shows the predictive accuracy of primacy and recency samples according to the *option-split* method. Random stopping clearly had no specific effect on recency and primacy. In line with our speculation, however, optional stopping had a pronounced effect, which lay in a similar range as the aggregate effect found for the three empirical studies. Thus, although previously neglected, optional stopping is a viable alternative account for recency in the sampling paradigm of decisions from experience.



*Figure 3.* Results of the random and optional stopping simulation: predictive accuracies of primacy and recency predictions. The first two pairs of bars show the results of the simulation; the last pair shows empirical results from three studies using the same set of as used in the simulation.

#### 4. Discussion

In this investigation, we evaluated and interpreted the existing evidence for recency in the sampling paradigm of decisions from experience. To this end, we first identified previously reported findings and evaluated their implications for the two established explanations of recency: limited memory capacity (Kareev, 2000; Kareev, Lieberman, & Lev, 1997) and step-by-step updating (Hogarth & Einhorn, 1992; March, 1996). Although the overall evidence was mixed, the type of sampling (i.e., self-terminated vs. fixed) appeared to play an important role. Specifically, individual studies indicated that recency occurs only when the decision to terminate search is made freely by participants themselves. To test whether this represents a general pattern, we compiled all available data obtained using the sampling paradigm in its three variants—free (self-terminated), matched, and fixed—and tested for order effects. The result was a clear recency effect for free and matched sampling, irrespective of which of three methods was applied to test for recency, and no recency effect for fixed sampling.

Together with the additional limitations highlighted in our literature review—in particular, the absence of an association between recency and working memory (Rakow et al., 2008) and the finding that frequency judgments are, in general, very accurate (Camilleri & Newell, 2009a; Lejarraaga, 2010; Ungemach et al., 2009; see also Zacks & Hasher, 2002)—our results cast considerable doubt on the role of memory limitations for order effects in the sampling paradigm. Specifically, if memory limitations played a role in producing recency, one would expect signs of recency in the fixed sampling variants, especially given that some of the fixed data sets required the participant to sample multiple times (specifically, Camilleri & Newell, 2011b; Hau et al., 2008; Ungemach et al., 2009).

The step-by-step updating account is more consistent with selective recency for self-terminated sampling. Drawing on the work of Hogarth and Einhorn (1992), we argued here and elsewhere (Wulff & Pachur, 2014) that the requirement to terminate search likely elicits information processes akin to step-by-step belief updating, whereas fixed sampling does not necessarily do so. Because step-by-step updating is believed to lead to recency under most conditions, whereas its counterpart, end-of-sequence processing, does not, this provides an attractive explanation that connects the fields of belief or impression updating with decisions from experience.

Equally consistent, however, with the finding of selective recency is the optional stopping account introduced here. In its core, optional stopping proposes that recency occurs simply because participants stop sampling when the difference between the options is large. It represents a conceptually very different explanation to the other two (limited memory capacity and step-by-step belief updating) both of which explain recency on the level of how a given sequence is processed. Optional stopping requires no such assumptions. In fact, we would claim that any valuation mechanism, be it limited by memory or akin to a CPT-like evaluation, can lead to recency when search is terminated based on the same mechanism that is terminally used to make a decision. To support this claim, we repeated the simulation using utilities derived from CPT (based on the formulation and parameter values observed by Tversky & Kahneman, 1992) and found identical results.

Returning to the debate on whether decisions from experience, as exemplified by the sampling paradigm, and decisions from description require independent sets of theories, it thus seems that our investigation may not only provide answers, but also raises new questions. Clearly, recency is a robust phenomenon in the sampling paradigm. However, whether it implies that information processing takes different paths in decisions from

experience versus decision from description due to the format in which the information is available remains an open question.

If optional stopping did prove to be the most plausible explanation of recency in the sampling paradigm, a new level of differences between decisions from experience and decisions from description would be introduced. One additional consequence of optional stopping must be that the experience of options is considerably distorted. Indeed, although we did not implement it in exactly that way, optional stopping in the sampling paradigm is essentially identical to maximizing a particular test statistic by optionally stopping in data collection (Berger & Berry, 1988). As optional stopping in data collection results in more extreme effect sizes, the differences between the experiences of the two options will also become more extreme. Because the possible outcomes are predetermined, more extreme observed differences must imply systematic deviations between the observed probabilities and the true probabilities.

Unfortunately, optional stopping in the sampling paradigm is an elusive beast. We made several attempts to establish optional stopping with the present data; however, we were not able to arrive at conclusive results. One important reason for this is that optional stopping most likely depends on the preference structure of a given person, such as the degree of risk or loss aversion (Kahneman & Tversky, 1979; Lopes & Oden, 1999). For instance, someone who is extremely averse to risks may take many more samples after a desirable rare event in order to reduce its perceived likelihood of occurrence. Someone with a greater tendency to seek risks, however, may stop at that point or take only a few further samples. Elsewhere, we present indirect evidence for exactly this behavioral pattern (Wulff et al., 2014b). This kind of preference-dependent optional stopping complicates matters, because it is impossible to tell whether a person stopped at a suitable point in time—where the options appeared sufficiently

distinct in respect to sample size—without knowing how that person evaluates the available options.

Additional analyses and, probably, additional data are thus needed to determine whether optional stopping or step-by-step belief updating provides the more plausible account of recency in the sampling paradigm. The two explanations are currently on a par, as both are equally consistent with the pattern of results presented here, but both are also purely hypothetical at the present point. One potential avenue to gain additional insights would be to analyze the relationship of switching patterns and recency. Rakow et al. (2008) have made a compelling argument that those who switch frequently between options engage in more step-by-step processing. If future studies confirm such a relationship between recency and switching, it could be regarded as evidence in favor of step-by-step processing. With respect to optional stopping, future studies should attempt to measure the constituents of preference (e.g., risk and loss aversion) independently of behavior shown in the sampling paradigm. This would make it possible to evaluate whether participants are more likely to terminate search when, in their eyes, the options are more distinct. Moreover, it needs be considered that the exact nature of optional stopping may differ from the one implemented here. It is also possible to approach the process from the opposite angle by assuming that participants start out with a fixed sample size in mind and only decide to collect additional samples if they deem it necessary. Thus, optional stopping may in fact be “optional sampling.” Although generally consistent with our implementation of optional stopping, in which the likelihood of terminating search increased monotonically with sample size, optional sampling would be more discrete and harder to detect. Generally, we consider it highly important to address the possibility of optional stopping in future investigations using the sampling paradigm. For instance, another possible consequence of optional stopping or optional sampling might be

not only that aggregate properties of the options are falsely reflected in experience, but also that particular samples occur at nonrandom points in the sequence. Starting points for the development of models of optional stopping are given by recent research on self-directed learning (Gureckis & Markant, 2012; Markant & Gureckis, 2014) and hypothesis generation (Jahn & Braatz, 2014; Lange, Thomas, & Davelaar, 2012; Thomas, Dougherty, Sprenger, & Harbison, 2008).

Last but not least, at least one limitation of this study warrants consideration. Although all studies included in the meta-analysis applied nearly identical methodologies, they used different manipulations and may also have differed in other relevant aspects (e.g., sample population, number of problems, tasks completed before introduction of the sampling paradigm). Importantly, not only was it impossible to balance the manipulations across the sampling variants, but the amount of data for free, matched, and fixed sampling also differed considerably (see Table 2). Both aspects could have influenced the pattern of results. Critically, however, the three studies with high fixed samples sizes did not involve a meaningful manipulation, but still failed to produce recency for fixed sampling (Camilleri & Newell, 2011b; Hau et al., 2008; Ungemach et al., 2009). While we see this as strong evidence against recency under fixed sampling, more data would be helpful, as would attempts to align the findings of the studies with respect to the manipulations used.

In sum, we have demonstrated that recency is a real and robust phenomenon in the sampling paradigm when, and only when, participants terminate search actively. We have also demonstrated that, alongside step-by-step belief updating, optional stopping is equally capable of explaining the current pattern of results, as both are consistent with recency in free sampling, but not in fixed sampling. We leave it future studies to judge which of the two provides the more viable explanation of recency. For the debate on the description–experience

gap, our investigation highlights that decisions from experience, as implemented by the sampling paradigm, call for at least one additional conceptual ingredient beside limited information search. Whether or not this ingredient is an altered style of information processing due to differences in the formats of experience and description, however, depends on which mechanism proves to be the best explanation for recency. Finally, and perhaps most importantly, our investigation indicates that the role of memory in the sampling paradigm needs to be revised. The fact that the predominant computational models for the sampling paradigm—the primed sampler model (Erev et al., 2010) and the Instance-Based Learning model (IBL, Gonzalez & Dutt, 2011)—interpret recency as a function of memory highlights that some rethinking is required in the literature on the sampling paradigm of decisions from experience.



### References

\*Studies included in the meta analysis

- Anderson, N. H., & Hovland, C. I. (1957). In C. I. Hovland (Ed.), *The order of presentation in persuasion* (pp. 158-169). New Haven, CT: Yale Univ. Press.
- Atkinson, R. C., & Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. *Psychology of Learning and Motivation*, 2, 89–195.
- Barron, G., & Erev, I. (2003). Small feedback-based decisions and their limited correspondence to description-based decisions. *Journal of Behavioral Decision Making*, 16, 215–233.
- Berger, J. O., & Berry, D. A. (1988). Statistical analysis and the illusion of objectivity. *American Scientist*, 76, 159–165.
- \*Camilleri, A. R., & Newell, B. R. (2009a). The role of representation in experience-based choice. *Judgment and Decision Making*, 4(7), 518–529.
- \*Camilleri, A. R., & Newell, B. R. (2009b). Within-subject preference reversals in description-and experience-based choice. *Cog Sci Society*, 31, 449–454.
- \*Camilleri, A. R., & Newell, B. R. (2011a). Description- and experience-based choice: Does equivalent information equal equivalent choice? *Acta Psychologica*, 136(3), 276–284.
- \*Camilleri, A. R., & Newell, B. R. (2011b). When and why rare events are underweighted: A direct comparison of the sampling, partial feedback, full feedback and description choice paradigms. *Psychonomic Bulletin & Review*, 18(2), 377–384.
- Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, 24, 87–114.  
doi:10.1017/S0140525X01003922
- Erev, I., Ert, E., Roth, A. E., Haruvy, E., Herzog, S. M., Hau, R., . . . Lebiere, C. (2010). A choice prediction competition: Choices from experience and from description. *Journal*

*of Behavioral Decision Making*, 23, 15–47.

Fox, C. R., & Hadar, L. (2006). “Decisions from experience” = sampling error plus prospect theory: Reconsidering Hertwig, Barron, Weber & Erev (2004). *Judgment and Decision Making Journal*, 1(2), 159–161.

\*Frey, R., Hertwig, R., & Rieskamp, J. (2014). Fear shapes information acquisition in decisions from experience. *Cognition*, 132(1), 90–99.

Glöckner, A., Fiedler, S., Hochman, G., Ayal, S., & Hilbig, B. E. (2012). Processing differences between descriptions and experience: a comparative analysis using eye-tracking and physiological measures. *Frontiers in Psychology*, 3: 173.

Gonzalez, C., & Dutt, V. (2011). Instance-based learning: Integrating sampling and repeated decisions from experience. *Psychological Review*, 118(4), 523–551.

Gonzalez, C., & Dutt, V. (2012). Refuting data aggregation arguments and how the instance-based learning model stands criticism: A reply to Hills and Hertwig (2012). *Psychological Review*, 119, 893–898.

Gureckis, T. M., & Markant, D. B. (2012). Self-directed learning: A cognitive and computational perspective. *Perspectives on Psychological Science*, 7(5), 464–481.

\*Hadar, L., & Fox, C. R. (2009). Information asymmetry in decision from description versus decision from experience. *Judgment and Decision Making*, 4(4), 317–325.

\*Hau, R., Pleskac, T. J., Kiefer, J., & Hertwig, R. (2008). The description–experience gap in risky choice: The role of sample size and experienced probabilities. *Journal of Behavioral Decision Making*, 21, 493–518.

\*Hertwig, R., Barron, G., Weber, E. U., & Erev, I. (2004). Decisions from experience and the effect of rare events in risky choice. *Psychological Science*, 15(8), 534–539.

Hertwig, R., Barron, G., Weber, E. U., & Erev, I. (2006). The role of information sampling in

- risky choice. In K. Fiedler & P. Juslin (Eds.), *Information sampling and adaptive cognition* (pp. 72–91). New York, NY: Cambridge University Press.
- Hertwig, R., & Erev, I. (2009). The description–experience gap in risky choice. *Trends in cognitive sciences*, 13(12), 517–523.
- \*Hertwig, R., & Pleskac, T. J. (2010). Decisions from experience: Why small samples? *Cognition*, 115(2), 225–237.
- \*Hills, T. T., Noguchi, T., & Gibbert, M. (2013). Information overload or search-amplified risk? Set size and order effects on decisions from experience. *Psychonomic Bulletin & Review*, 20(5), 1023–1031.
- Hogarth, R. M., & Einhorn, H. J. (1992). Order effects in belief updating: The belief-adjustment model. *Cognitive Psychology*, 24(1), 1–55.
- Jahn, G., & Braatz, J. (2014). Memory indexing of sequential symptom processing in diagnostic reasoning. *Cognitive Psychology*, 68, 59–97.
- Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica: Journal of the Econometric Society*, 263–291.
- Kareev, Y. (2000). Seven (indeed, plus or minus two) and the detection of correlations. *Psychological Review*, 107(2), 397–402.
- Kareev, Y., Lieberman, I., & Lev, M. (1997). Through a narrow window: Sample size and the perception of correlation. *Journal of Experimental Psychology: General*, 126(3), 278–287.
- Lange, N. D., Thomas, R. P., & Davelaar, E. J. (2012). Temporal dynamics of hypothesis generation: The influences of data serial order, data consistency, and elicitation timing. *Frontiers in Psychology*, 3: 215.
- \*Lejarraga, T. (2010) When experience is better than description: Time delays and

- complexity. *Journal of Behavioral Decision Making*, 23(1), 100–116.
- Lejarraga, T., Hertwig, R., & Gonzalez, C. (2012). How choice ecology influences search in decisions from experience. *Cognition*, 124, 334–342.
- Lopes, L. L., & Oden, G. C. (1999). The role of aspiration level in risky choice: A comparison of cumulative prospect theory and SP/A theory. *Journal of Mathematical Psychology*, 43(2), 286–313.
- March, J. G. (1996). Learning to be risk averse. *Psychological Review*, 103, 309–319.
- Markant, D. B., & Gureckis, T. M. (2014). Is it better to select or to receive? Learning via active and passive hypothesis testing. *Journal of Experimental Psychology: General*, 143(1), 94–122.
- \*Mehlhorn, K., Ben-Asher, N., Dutt, V., & Gonzalez, C. (2014). Observed variability and values matter: Toward a better understanding of information search and decisions from experience. *Journal of Behavioral Decision Making*, 27, 328–339.
- Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, 63, 81–97.
- Murdock Jr., B. B. (1962). The serial position effect of free recall. *Journal of Experimental Psychology*, 64(5), 482–488.
- Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running experiments on Amazon Mechanical Turk. *Judgment and Decision making*, 5(5), 411–419.
- \*Phillips, N. D., Hertwig, R., Kareev, Y., & Avrahami, J. (2014). Rivals in the dark: How competition influences search in decisions under uncertainty. *Cognition*, 133(1), 104–119.
- \*Rakow, T., Demes, K. A., & Newell, B. R. (2008). Biased samples not mode of presentation: Re-examining the apparent underweighting of rare events in experience-

based choice. *Organizational Behavior and Human Decision Processes*, 106(2), 168–179.

\*Rakow, T., & Rahim, S. B. (2010). Developmental insights into experience-based decision making. *Journal of Behavioral Decision Making*, 23(1), 69–82.

Rakow, T., & Newell, B. R. (2010). Degrees of uncertainty: An overview and framework for future research on experience-based choice. *Journal of Behavioral Decision Making*, 23(1), 1–14.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests*. Chicago: University of Chicago Press.

Thomas, R. P., Dougherty, M. R., Sprenger, A. M., & Harbison, J. (2008). Diagnostic hypothesis generation and human judgment. *Psychological Review*, 115(1), 155.

Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, 5, 297–323.

\*Ungemach, C., Chater, N., & Stewart, N. (2009). Are probabilities overweighted or underweighted when rare outcomes are experienced (rarely)? *Psychological Science*, 20(4), 473–479.

Weber, E. U., Shafir, S., & Blais, A. R. (2004). Predicting risk sensitivity in humans and lower animals: Risk as variance or coefficient of variation. *Psychological Review*, 111(2), 430–445.

\*Wulff, D., & Hertwig, R. (2014). Unpublished study.

\*Wulff, D. U., Hills, T. T., & Hertwig R. (2014) How short- and long-run aspirations impact search and choice in decisions from experience. Submitted manuscript.

Wulff, D. U., Hills, T. T., & Hertwig, R. (2014). Online product reviews and the description–experience gap. *Journal of Behavioral Decision Making*. Advance online publication.

doi:10.1002/bdm.1841

Wulff, D., & Pachur, T. (2014). Modeling valuations from experience. Manuscript in preparation.

Zacks, R.T., & Hasher, L. (2002). Frequency processing: A twenty-five year perspective. In P. Sedlmeier & T. Betsch (Eds.), *ETC frequency processing and cognition* (pp. 21–36). New York, NY: Oxford University Press.

## Appendix A

Table A1. *Results Retrieved from the Literature*

Paper	Method	Result (as reported; 1 <sup>st</sup> vs. 2 <sup>nd</sup> half)	Interpretation (as reported)	Sampling type	Notes
Hertwig et al. (2004)	Option-split	59% vs. 75%	<b>Recency</b>	Free	Same data (VUM reported in Hertwig et al., 2006)
	Value-updating model	$\phi = .29$	<b>Recency</b>	Free	
Rakow et al. (2008)	Option-split	66% vs. 76%	<b>Recency</b>	Free	
Hau et al. (2008)	Option-split	58% vs. 60%	No effect	Free	
Ungemach et al. (2009)	Option-split	65% vs. 59%	No effect	Free	
	Option-split	42% vs. 48%	No effect	Matched	
Camilleri & Newell (2009a)	Option-split	56% vs. 61%	No effect	Matched	
Rakow & Rahim (2010)	Option-split	69% vs. 55%	<b>Primacy</b>	Fixed	Adults, Exp1
	Option-split	64% vs. 64%	No effect	Fixed	Adolescents, Exp2
	Option-split	64% vs. 56%	<b>Primacy</b>	Fixed	Adolescents, Exp3
	Option-split	58% vs. 57%	No effect	Fixed	Young adolescents, Exp3
	Option-split	50% vs. 56%	<b>Recency</b>	Fixed	Children, Exp1
	Option-split	60% vs. 66%	<b>Recency</b>	Fixed	Children, Exp2
Camilleri & Newell (2011a)	Problem-split	63% vs. 49%	<b>Recency</b>	Fixed	Same data
	Option-split	—	No effect	Fixed	
Camilleri & Newell (2011b)	Option-split	39% vs. 65%	<b>Recency</b>	Free	
	Option-split	47% vs. 54%	No effect	Matched	Same data (Exp1)
	Problem-split	—	No effect		
	Option-split	57% vs. 51%	No effect	Matched	Same data (Exp2)
	Problem-split	—	No effect		
Wulff et al. (2014)	Switch-split	26% vs. 74%	<b>Recency</b>	Free	Only discriminating trials

Table A2. *Units of Data Sets used in the Analysis Ordered by Year of Publication*

Paper	Short	Unit	N	Problems	Certain	Problems/ Participant	Sampling type	Median sample size	Notes
Hertwig et al. (2004)	HBWE04	1	50	4G, 2L	4	3	Free	17.2	
Hau et al. (2008)	HPKH	1	42	4G, 2L	4	6	Free	10.6	Study 1
		2	39	4G, 2L	4	6	Free	33.5	Study 2, High incentives
		3	40	4G, 2L	4	6	Fixed	100	Study 3
Rakow et al. (2008)	RDN08	1	80	9G, 3L	9	6	Free	15.3	
Camilleri & Newell (2009a)	CN09a	1	80	4G, 4L	8	2	Matched	25.2	Frequency judgment after choice
		2	80	4G, 4L	8	2	Matched	27.2	Frequency judgment before choice
Camilleri & Newell (2009b)	CN09b	1	20	7G, 3L	10	10	Free	9.1	Experience first
		2	20	7G, 3L	10	10	Free	11.5	Description first
Hadar & Fox (2009)	HF09	1	23	2G, 1L	2	3	Fixed	20	Complete information, sampled amounts
		2	31	2G, 1L	2	3	Fixed	20	Incomplete information, sampled shapes
		3	30	2G, 1L	2	3	Fixed	20	Incomplete information, sampled amounts
		4	27	2G, 1L	2	3	Fixed	20	Complete information, sampled shapes
Ungemach et al. (2009)	UCS09	1	25	4G, 2L	4	6	Free	21.2	
		2	25	4G, 2L	4	6	Fixed	21.2	Yoked design
		3	197	4G, 2L	4	1	Fixed	80	
Rakow & Rahim (2010)	RR10	1	26	4G	4	4	Fixed	20	Study 1, 5–6 years old
		2	25	4G	4	4	Fixed	20	Study 1, adults
		3	38	6G	6	6	Fixed	20	Study 2, 5–6 years old
		4	37	6G	6	6	Fixed	20	Study 2, 16–17 years old, experience first



Paper	Short	Unit	<i>N</i>	Problems	Certain	Problems/ Participant	Sampling type	Median sample size	Notes
Rakow & Rahim (2010) continued	RR10	5	40	6G	6	6	Fixed	20	Study 2, 16–17 years old, description first
		6	19	2G, 4M	6	6	Fixed	20	Study 3, 12–13 years old, experience first
		7	20	2G, 4M	6	6	Fixed	20	Study 2, 16–17 years old, experience first
		8	17	2G, 4M	6	6	Fixed	20	Study 3, 12–13 years old, description first
		9	17	2G, 4M	6	6	Fixed	20	Study 2, 16–17 years old, description first
Hertwig & Pleskac (2010)	HP10	1	88	8G, 4L	8	12	Free	10.9	
Lejarraga (2010)	L10	1	85	4G, 3L	3	3	Free	35	Exp2, self-selected
Erev et al. (2010)	TPT10	1	39	20G, 20L, 20M	60	30	Free	10.5	Technion prediction competition, estimation
		2	40	20G, 20L, 20M	60	30	Free	13.7	Technion prediction competition, estimation
Camilleri & Newell (2011a)	CN11a	1	40	2G, 2L	4	4	Fixed	100	
Camilleri & Newell (2011b)	CN11b	1	31	7G, 3L	10	10	Matched	10.1	Pseudo-random
		2	35	7G, 3L	10	10	Matched	11.6	Pseudo-random
		3	36	5G, 3L	8	8	Fixed	20	
Glöckner, Fiedler, Hochman, Ayal, & Hilbig (2012)	GFHAH12	1	22	37G	0	37	Free	33.5	Eye-tracker, target problems
		2	22	22G	0	22	Free	28.8	Eye-tracker, filler problems
Hills, Noguchi, & Gibbert (2013)	HNG13	1	32	1G	0	1	Free	4.5	One-to-many
		2	32	1G	0	1	Free	5.5	Many-to-one
Frey et al. (2014)	FHR14	1	27	5G, 4L	5	9	Free	33.1	Mood induction: Happy
		2	28	5G, 4L	5	9	Free	30.6	Mood induction: Sad
		3	29	5G, 4L	5	9	Free	50.4	Mood induction: Fearful
		4	28	5G, 4L	5	9	Free	34.1	Mood induction: Angry
		5	23	2G, 2L	4	4	Free	51.2	Field: Dental surgeon

*Note.* Short: Abbreviation of the article used in Figure 2. *N*: Number of participants in the unit. Problems: Number of problems using gain (G), loss (L), and mixed (M) decision problems. Certain: Number of problems containing a sure event option.

Table A2 continued. Data sets used in analysis.

Paper	Short	Unit	<i>N</i>	Problems	Certain	Problems/ Participant	Sampling type	Median sample size	Notes
Frey et al. (2014) continued	FHR14	6	26	2G, 2L	4	4	Free	4.9	Field: Comedy show
Phillips, Hertwig, Kareev, & Avrahami (2014)	PHKA14	1	36	21M	0	5	Free	20.9	Social competition Mechanical Turk
		2	142	21M	0	3	Free	1.7	
Mehlhorn, Ben-Asher, Dutt, & Gonzalez (2014)	MBDG14	1	294	8G, 8L	16	2	Free	3.5	
Wulff, Hills, & Hertwig (2014)	WHH14	1	41	16G	9	16	Free	19	Single-play framing
		2	42	16G	9	16	Free	24.9	Multi-play framing
		3	41	16G	9	16	Free	19.6	
Wulff & Hertwig (2014)	WH99	1	59	4G, 2L	4	6	Free	15.3	Mechanical Turk, experience first
		2	78	4G, 2L	4	6	Free	10.7	Mechanical Turk, description first
		3	41	4G, 2L	4	6	Free	14.8	Mechanical Turk, experience first, US only
		4	40	4G, 2L	4	6	Free	8.7	Mechanical Turk, description first, US only

*Note.* Short: Abbreviation of the article used in Figure 2. *N*: Number of participants in the unit. Problems: Number of problems using gain (G), loss (L), and mixed (M) decision problems. Certain: Number of problems containing a sure event option.

## Appendix B

The three methods used to test for recency were implemented as follows (see Fig. 1):

*Option-Split.* Each option was split in half with respect to the number of samples drawn from it. For example, if six samples were drawn from one option, the first three samples were used for the primacy prediction and the second three for the recency prediction. If the number of samples from an option was uneven, the middle sample was used for both predictions. The primacy prediction was determined by calculating the means for the primacy subsamples of each option and selecting the option with the higher mean. If the means for both options were identical, no prediction was made and the case was excluded from the analysis. The recency prediction was determined analogously.

*Problem-Split.* Each problem was split in half with respect to the number of samples drawn in total. For example, if 12 samples were drawn in total, the first six samples were used for the primacy prediction and the second six for the recency prediction. If the number of samples from an option was uneven, the middle sample was used for both predictions. The primacy prediction was determined by calculating the means for the primacy subsamples of each option and selecting the option with the higher mean. If the means for both options were identical or if samples were available for only one option, no prediction was made and the case was excluded from the analysis. The recency prediction was determined analogously.

*Switch-Split.* For each problem, the transitions from one option to the other (switches) were determined. For the primacy prediction, all samples around the first switch were used—that is, all samples before the second switch. For the recency prediction, all samples around the last switch were used—that is, all samples after the second-to-last switch. In cases where the primacy and recency subsamples overlapped, which occurred when participant switched exactly once or twice, the samples were used for both predictions. The primacy prediction

was determined by calculating the means for the primacy subsamples of each option and selecting the option with the higher mean. If the means for both options were identical, no prediction was made and the case was excluded from the analysis. The recency prediction was determined analogously.

# Modeling Valuations From Experience

Dirk U. Wulff and Thorsten Pachur

Max Planck Institute for Human Development, Berlin

## Author Note

Dirk U. Wulff and Thorsten Pachur, Center for Adaptive Rationality, Max Planck Institute for Human Development, Berlin, Germany.

We are grateful to Susannah Goss for editing the manuscript

Correspondence concerning this article should be addressed to Dirk U. Wulff, Center for Adaptive Rationality (ARC), Max Planck Institute for Human Development, Lentzeallee 94, 14195 Berlin, Germany. Phone: +49 30 82406 475. E-mail: [wulff@mpib-berlin.mpg.de](mailto:wulff@mpib-berlin.mpg.de)

## Abstract

What are the cognitive mechanisms underlying valuations based on sequentially experienced samples of an option's possible outcomes? Ashby and Rakow (2014) have proposed a *sliding window model* (SWIM), according to which people's valuations represent the average of a limited sample of recent experiences (the size of which is estimated by the model) formed after sampling has been terminated (i.e., an end-of-sequence process). Ashby and Rakow present results from which they conclude that the SWIM performs well compared to alternative models based on model selection criteria (value-updating model, summary model). Further, they report that the individual window sizes estimated by the SWIM correlate with a measure of working memory capacity. We highlight several problematic issues with the conclusions drawn by Ashby and Rakow. In a reanalysis of their data, we find no clear evidence in support of any of the models tested, but with a slight advantage for the summary model. Further, we demonstrate that individual differences in the window size estimated by the SWIM can reflect differences in judgment noise. Model recovery analyses reveal that the flexibility of the models tested by Ashby and Rakow depends on a complex interplay of sample size and noise, precluding unequivocal conclusions regarding the underlying mechanism on the basis of the data presented. We discuss several approaches to improve model comparisons in valuations from experience.

*Keywords:* valuations from experience, active sampling, cognitive modeling, model complexity, monetary gambles.

## Modeling Valuations From Experience

When a valuation of an object is formed on the basis of sequential experiences, not all experiences with the object necessarily contribute equally to the valuation. Instead, more recent experiences tend to have a stronger impact than do less recent ones (e.g., Hogarth & Einhorn, 1992). How can experience-based valuations, in particular potential recency effects, best be modeled? A common assumption is that more distant experiences have a gradually decreasing influence, in line with the idea that memory traces decay with time (Ebbinghaus, 1885; Rubin & Wenzel, 1996). A prominent instantiation of this notion in decision research is the *value-updating model* (VUM; Hertwig, Barron, Weber, & Erev, 2006), according to which a valuation  $v$  after experiencing  $n$  outcomes is determined as follows:

$$v_n = \left[1 - \left(\frac{1}{n}\right)^\varphi\right] v_{n-1} + \left(\frac{1}{n}\right)^\varphi x_n. \quad (1)$$

The parameter  $\varphi$  either gives more weight to earlier samples ( $\varphi > 1$ , *primacy*) or to later samples ( $\varphi < 1$ , *recency*), or weights all samples equally ( $\varphi = 1$ ; for related models, see Yechiam & Busemeyer, 2006).<sup>1</sup>

Ashby and Rakow (2014) recently proposed an intriguing alternative way of modeling a stronger influence of more recent experiences. Rather than assuming a gradually decreasing impact, they postulated an all-or-nothing mechanism that considers all experiences in a recent, typically limited window and that can exclude more distant experiences. Specifically, the *sliding window model* (SWIM) predicts that a valuation  $v_n$  is formed by averaging  $\zeta$  out of  $n$  total experiences  $x_i$ :

---

<sup>1</sup> Consistent with the original specification in Hertwig et al. (2006), Ashby and Rakow (2014) applied a nonlinear transformation to the experienced outcomes ( $x_n^{.88}$  instead of  $x_n^1$ ). As we do not include this peripheral aspect of the model in our analyses for matters of model comparability, we display the model with linear outcome weighting.

$$v_n = \frac{1}{\zeta} \sum_{i=1+n-\zeta}^n x_i \quad (2)$$

If the size of the window  $\zeta$  is smaller than the total number of experiences  $n$ , the SWIM implements recency effects, by assuming that some (namely,  $n$  minus  $\zeta$ ) of the earliest experiences are completely dropped from consideration; all experiences within the window contribute equally to the valuation. This account of recency is in line with models of working memory that posit a fixed, limited storage capacity, and where an item is currently either activated in memory or not (e.g., Cowan, 2001).

Ashby and Rakow (2014) highlighted that the SWIM should be interpreted as an *end-of-sequence mechanism*, where an evaluation is formed only after the sampling process has been terminated. This deviates from the assumption in the VUM (and other models; e.g., Bush & Mosteller, 1955; Hogarth & Einhorn, 1992; March, 1996) of a *step-by-step mechanism*, where an evaluation is continually formed and updated on-line during the sampling process. (Hastie and Park [1986] made a related distinction between memory-based and on-line judgments.)

Ashby and Rakow (2014) reported two empirical studies in which they pitted the SWIM against the VUM as well as the *summary model* (SUM; Hills & Hertwig, 2010; Wulff, Hills, & Hertwig, 2014), which calculates an average across all experiences (thus assuming no recency). They concluded “that for many individuals not all information is used and that the amount of information integrated is, in part, related to individual differences in cognitive abilities such as memory span” (p. 1160). This conclusion is based on several findings. First, the SWIM showed a better average fit on the Akaike Information Criterion (AIC; Akaike, 1973). Second, the window size estimated for individual participants using the SWIM was consistently smaller than the average number of samples the participant had drawn. Third,



there was a higher correlation between sample size and response times for people better fit by the SWIM than for those better fit by the VUM or SUM, in line with the assumption of an end-of-sequence process (which predicts that the more experiences that can be retrieved, the longer the response should take). Fourth, for participants better fit by the SWIM, the estimated window size was positively related to working memory capacity, as measured by a digit span task.

The SWIM represents an attractive addition to the growing literature on models of experienced-based judgment and decision making (e.g., Hertwig, *in press*; Hertwig & Erev, 2009), and the empirical findings presented by Ashby and Rakow (2014) are intriguing. Moreover, the proposed all-or-nothing nature of the consideration of sampled outcomes has nice conceptual similarities with the limited capacity assumption in prominent conceptions of working memory (Baddeley, 2012), and thus promises to reinforce the link between decision making and working memory research. In this paper, however, we argue that the evidence presented may not yet warrant the conclusion that experience-based valuations are based on an all-or-nothing, end-of-sequence evaluation process, as assumed in the SWIM.

In the first part, we critically reevaluate the model comparison conducted by Ashby and Rakow (2014), finding that if the data do support one particular model, it is the SUM (which assumes no recency at all) rather than the SWIM. We then show that under a process where all experiences contribute equally to a valuation, different amounts of noise are reflected in SWIM as different amounts of recency; it is therefore unclear whether individual differences in window size estimated by the SWIM indeed reflect the amount of information considered in the judgment—or individual differences in unsystematic responding. This possibility may complicate the interpretation of Ashby and Rakow that correlations between the SWIM window size estimates and working memory capacity support the specific

processes assumed by the model. Finally, we scrutinize the assumption of an end-of-sequence process on a conceptual level, pointing out that the requirement to conduct and terminate search actively must elicit some form of step-by-step process, which is at odds with the idea of a pure end-of-sequence process.

In the second part, we examine valuations from experience and the proposed models with respect to two potential explanations for the inconclusive pattern of results presented in the first part: a mixture of different processes and low model discriminability. Finding evidence for both, we conclude that it is practically impossible to identify the underlying recency process when comparing the SUM, VUM, and SWIM on the basis of the design used by Ashby and Rakow (2014). Finally, we derive implications and suggestions for the study of valuations from experience that are relevant for the success of future investigations using this paradigm, but also for the field of experience-based decision making in general (see Hertwig & Erev, 2009).

### **Reanalysis of Ashby and Rakow (2014)**

#### **Model Comparison**

A key basis for Ashby and Rakow's (2014) conclusion regarding the viability of the SWIM was its performance in a model comparison pitting it against the VUM and the SUM. The authors used two popular measures to evaluate the three models: the Bayesian Information Criterion (BIC; Schwarz, 1978) and the AIC. Both indices penalize for model complexity based on the number of free parameters.<sup>2</sup> Ashby and Rakow considered two

---

<sup>2</sup> BIC approximates the marginal likelihood of the data given a specific model and AIC approximates the Kullback-Leibler divergence between the true and the evaluated model. Which of the two measures is to be preferred is debated (for overviews, see Burnham & Anderson, 2002, 2004; Lewandowsky & Farrell, 2010; Vrieze, 2012).

aspects of model performance: the number of participants best accounted for by each model (according to BIC and AIC), and each model's median and average (across participants) BIC and AIC. According to BIC, the VUM and the SUM performed best in Studies 1 and 2, respectively, in terms of the number of participants. According to AIC, SWIM accounted for the largest number of participants. However, statistical tests showed no significant differences in the percentages of best model fit for either BIC or AIC. Based on the average BIC and AIC values, SWIM emerged as the best model, although no attempt was made to substantiate these differences statistically.

There are several issues with using the results of the model comparison as conducted by Ashby and Rakow (2014) to draw conclusions regarding the viability of a mechanism as assumed in the SWIM. First, the SWIM assumes a linear value function for the outcomes, whereas the VUM, as implemented by Ashby and Rakow, assumes a nonlinear value function (i.e.,  $x^\alpha$  with  $\alpha = .88$ ). Moreover, the VUM allowed for primacy, whereas the SWIM did not. These differences complicate the recovery of the underlying recency mechanism because other, noncentral aspects may cause the model to perform well or poorly.

Second, the models were not fit with equal precision. This issue concerns the two-stage fitting procedure applied by Ashby and Rakow (2014). To simplify the estimation of parameters, they first determined the best fitting standard deviation of the noise distribution implemented to derive likelihoods, before then estimating the recency parameters and the maximum likelihood of the models. Importantly, this pre-estimation of the standard deviation was done using a model that equated to the SUM. This approach thus reduced the flexibility in the SWIM and VUM relative to the SUM, although all models were punished alike for estimation noise.

Third, the confinement of outcomes and valuations to the doubly bound interval between 0 and 4, as done in Ashby and Rakow (2014), strongly suggests the use of a truncated error distribution. However, Ashby and Rakow used an untruncated normal distribution to account for deviations from the models' predictions. The consequence of an untruncated error distribution is that predictions close to the boundaries (i.e., close to 0 and 4) receive generally less weight, as in such cases greater portions of the probability mass are cut off by the boundaries.<sup>3</sup>

To address these concerns, we reanalyzed the data of Ashby and Rakow (2014) using a more refined approach to model estimation and evaluation. Specifically, (1) we used a linear value function for all models, (2) we fitted two versions of the VUM, one allowing for recency only, which we refer to as VUMr ( $\phi = [0, 1]$ ), and one allowing for both recency and primacy ( $\phi = [0, \text{Inf}]$ ), (3) we estimated all model parameters simultaneously, (4) we used a combination of grid search and subsequent optimization using quasi-Newton minimization, (5) we used a truncated normal distribution to model noise and thus properly match the doubly bound valuation interval between 0 and 4, and finally (6) we used the  $AIC_c$ ,<sup>4</sup> which

---

<sup>3</sup> Furthermore, when we recalculated the models' AICs using the fits reported in the supplementary material, we obtained somewhat different results. Specifically, it appears that Ashby and Rakow (2014) subtracted the penalty term of the AIC instead of adding it, which strongly favors the two more complex models, VUM and SWIM. When this is corrected, the SWIM is no longer superior. Additionally, there were some minor inconsistencies in the BICs and likelihoods. A recent erratum should, however, have corrected these issues.

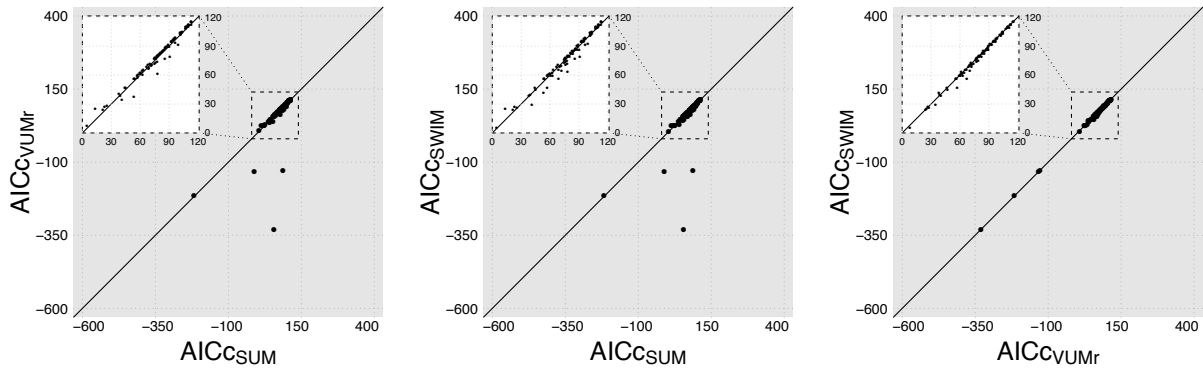
<sup>4</sup> The AIC is known to penalize insufficiently for model flexibility at small sample sizes. For this reason, Burnham and Anderson (2002) have recommended the use of an adjusted index, the  $AIC_c$ , which corrects for this bias (for large samples,  $AIC_c$  approaches AIC) and is defined as follows:

$$AIC_c = -2\ln(L) + 2k + \frac{2k(k+1)}{n-k-1}.$$

with  $L$  being the likelihood,  $k$  the number of parameters, and  $n$  the number of data points used to calculate the likelihood. Given that each participant in Ashby and Rakow's (2014) studies provided only a modest number of data points (valuations of 40 lotteries; moreover, for some

provides more accurate approximations of the Kullback-Leibler divergence for finite data (Burnham & Anderson, 2002) than does the AIC.

In Figure 1, we first plot pairwise model comparisons of individual participants'  $AIC_c$  values (results for BIC values are qualitatively the same). As can be seen, for three participants, the fit was much better for the VUM and SWIM than for the SUM, with  $AIC_c$  differences of 123, 217, and 387—far beyond the next biggest absolute difference of 18. We excluded these three participants from all following analyses (unless indicated otherwise).<sup>5</sup>



*Figure 1.* Scatter plots showing pairwise model comparisons of individual participants'  $AIC_c$  values. VUMr is the recency-only VUM with  $\phi$  constrained to the interval  $[0, 1]$ . The white square in the upper left shows the majority of participants on a more fine-grained scale between 0 and 120.

---

participants, trials were excluded), the  $AIC_c$  seems a more appropriate index of model performance in the present case.

<sup>5</sup> Further analyses indicated that the three participants excluded from the analysis provided as the final valuation the last observed sample in at least 95% of the lotteries (although these participants drew, on average, more than one sample before making a valuation). Thus, these participants are likely to have applied a qualitatively different strategy than assumed by the SUM, VUM, or SWIM, which provides further grounds for their exclusion.

Figure 1 also shows that the three models had very similar fits to the data.

Summarizing this, Table 1 shows the percentage of participants in Ashby and Rakow (2014; aggregated across both studies) best accounted for by each model (including ties), separately for BIC and AIC<sub>c</sub>. According to both BIC (which usually punishes more strictly than AIC<sub>c</sub> for the number of free parameters) and AIC<sub>c</sub>, most participants were best accounted for by the SUM. Table 1 also presents the mean and median BIC and AIC<sub>c</sub>. These aggregate measures essentially show no differences between the models.

Table 1. *Model Fits Aggregated over Study 1 and 2 of Ashby and Rakow (2014). Shown Are the Number of Participants Best Fit by Each Model and Model Evaluation Criterion as well as the Mean and Median Criterion Value for Each Model.*

	% best fit	BIC		% best fit	AIC <sub>c</sub>	
		<i>Mdn</i>	<i>Mean</i>		<i>Mdn</i>	<i>Mean</i>
<i>SUM</i>	75	83.1	76.5	62	82.2	75.9
<i>VUM</i>	-	85	78.1	-	82.5	76.3
<i>VUMr</i>	7	85.3	78.3	10	82.7	76.5
<i>SWIM</i>	18	83	77.3	27	81.2	75.6

*Note.* Calculation of the AIC<sub>c</sub> led to an infinite criterion value for one participant, who appeared to have provided valuations for only three lotteries. We excluded this participant from the AIC<sub>c</sub> results.

In sum, whereas Rakow and Ashby (2014) derived from their analysis support for the mechanism assumed by the SWIM, a reanalysis of their data using a more refined model comparison approach shows a rather different pattern, with model performance depending considerably on the measure used for model evaluation (proportion best fit vs. median or mean information criterion value). Overall, there is little evidence that one model consistently

outperforms another. Consistent with the original results and the fact that the BIC punishes more strictly for model complexity (Burnham & Anderson, 2004), SUM receives more support from the BIC than the AIC<sub>c</sub> values. Inconsistent with the conclusions drawn by Ashby and Rakow, if the data do support one particular model, it is the SUM, which considers all experiences in the sample, rather than the SWIM.

### **Interpreting Noise as Forgetting?**

Ashby and Rakow's (2014) conclusion that valuations from experience (sometimes) follow an all-or-nothing, end-of-sequence process that considers only a limited number of the outcomes experienced was also based on findings beyond the results of the model comparison. First, the window sizes estimated by the SWIM were, on average, smaller than the number of samples drawn. Second, the estimated window size was related to a measure of working memory capacity, and this relationship was present for participants better fit by the SWIM, but absent for participants better fit by the VUM. This finding seems to support the interpretation that participants better fit by the SWIM integrate only as many samples as their working memory size permits. Third, in Study 2, participants better fit by the SWIM exhibited a more strongly positive relationship between window size and response time than did the other participants, as would be expected if the elements within the window were processed at the end of the sequence.

Assuming that at least some participants relied on a mechanism akin to the SWIM and were identified as such, a key requirement for this set of findings is that the window size estimated by the SWIM veridically reflects the number of experiences on which the valuation is based. As a critical test of this assumption, we simulated data using a special case of the SWIM where the true window size matches the samples. This allowed us to directly test Hypotheses 3 and 4 in Ashby and Rakow (2014), which propose that the window size should

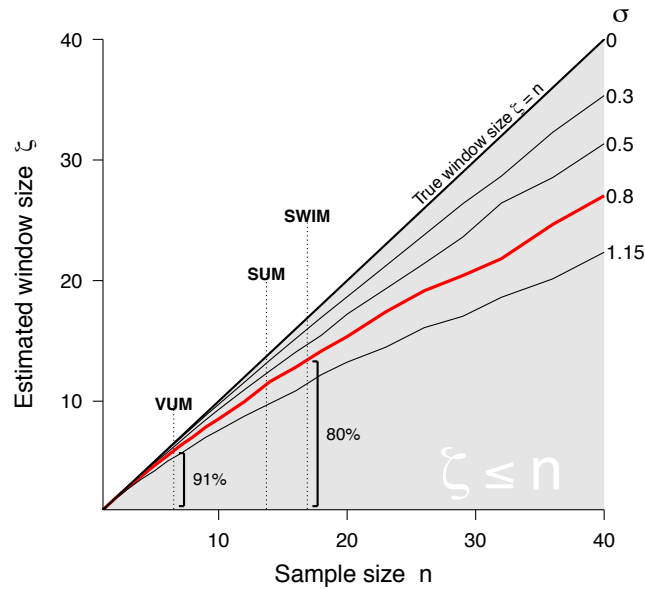
undershoot the actual sample size as a result of memory limitations, against an alternative hypothesis that this undershooting occurs solely due to noise in the data. Note that the SWIM must underestimate the window size (on average): point estimates from noisy data will always err to some extent, and with  $\zeta = n$  they can err only in the direction of smaller window sizes, not in the direction of larger window sizes.<sup>6</sup> To examine the extent of this noise-related undershooting, we examined how accurately the SWIM recovered the true window size (= sample size) over increasing levels of noise. Specifically, for each of various combinations of sample size and judgment noise, we generated 1,000 agents completing the task that Ashby and Rakow (2014) gave their participants. These agents first took a fixed number of samples for each of the 40 lotteries and then provided valuations based on a truncated normal distribution centered on the respective sample mean. Each agent's sample size and judgment noise were varied broadly, such that 95% of mean sample sizes and standard deviations found in the original data were covered (i.e., sample size = [1, 40] and  $\sigma$  = [.01, 1.15]). We then fit the SWIM to the each agent's data using an untruncated normal distribution (see Appendix A for details). We chose to use an untruncated normal distribution to establish comparability with the window sizes reported by Ashby and Rakow (2014).<sup>7</sup>

---

<sup>6</sup> If a model has a free parameter and this parameter is estimated from data containing noise, then the parameter estimate will most likely deviate from the actual parameter value. If this parameter can err in both directions (i.e., be larger or smaller than the true parameter value), then the estimation can still be unbiased. If, however, a parameter can err only in one direction, as is the case for the window size parameter with true  $\zeta = n$ , the estimator will on average not reflect the true parameter value, as it will be pulled exclusively to the one available direction. In the case of  $\zeta = n$ , this means that the estimate of  $\zeta$  is expected to be smaller than  $n$ .

<sup>7</sup> With a truncated normal distribution, the results became considerably more extreme. One explanation for the more moderate results for the untruncated normal distribution is this: The untruncated normal distribution punishes observations whose predictions lie close to the boundaries of 0 and 4, as greater portions of the error distribution for such values are cut off by the boundaries. Such extreme predictions are generally more likely for small window sizes than for large window sizes due to the law of large numbers and the specific problem structure used by Ashby and Rakow (2014). Thus, the untruncated normal distribution favors





*Figure 2.* Window size estimated by the SWIM for various levels of actual sample size and different levels of noise (with the SUM as the generating mechanism). The red line represents the mean noise level for the data of Ashby and Rakow (2014). The dashed lines labeled by the model acronyms represent the mean sample size of participants best fit by the respective model according to the classifications made by Ashby and Rakow.

Figure 2 plots the simulation results, showing the average estimated window size, as determined for data generated by the SUM, as a function of the sample size (i.e., the actual window size). The different lines represent different levels of noise (the standard deviation of the untruncated normal distribution). Note that the SUM is nested under the SWIM, because SWIM and SUM are identical for  $\zeta = n$ . If the window size estimated by the SWIM (i.e.,  $\zeta$ ) is a valid measure of the window size of experiences that underlie a valuation, the estimate

---

larger window sizes and as such tends to lead to more moderate predictions that lie further away from the boundaries.

should coincide with the actual sample size (i.e., the line should lie on the diagonal). As the figure reveals, however, noise leads to a systematic underestimation of the actual window size. Note that this result obtains even though the implementation of the SWIM allows noise to be captured. The red line in Figure 2 shows the mean noise level estimated for the data of Ashby and Rakow (2014). For instance, when the actual sample size is 14 (roughly the average sample size reported in Ashby and Rakow), the estimated window size is, on average, 11.6. Depending on the noise level, the average window size estimate when 14 observations contribute to the valuation can vary between 14 (noise = .01) and 9.8 (noise = 1.15).<sup>8</sup> The finding interpreted by Ashby and Rakow in support of the SWIM that the estimated window size undershoots the sample size is thus not necessarily an indicator that only a subset of experiences are considered.

The finding that noise can systematically influence the window size estimated by the SWIM thus calls for a reconsideration of the correlations between estimated window size and working memory capacity and response times reported by Ashby and Rakow (2014). Note that measures of working memory capacity can also reflect attentional control (Kane & Engle, 2003; Kane et al., 2007) and correlate with measures of general intelligence (e.g., Ackerman, Beier, & Boyle, 2005). Therefore, the reported correlations with working memory capacity may simply be due to the fact that people with lower cognitive capacities have a tendency to respond less systematically.

---

<sup>8</sup> In an additional analysis, we determined the estimated window size from our simulation for the individual observed sample size and noise levels—that is, we matched each participant in Studies 1 and 2 to the conditions in our simulation that were closest in terms of sample size and noise level. The average simulated window size of 9.5 is very close to the reported estimated window size across both studies (window size of 9). Hence, individual differences in window size estimated by SWIM may actually reflect individual differences in the level of noise, rather than (or in addition to) actual forgetting.

Finally, if window sizes estimated by the SWIM are not necessarily a valid measure of actual information use, how is it possible to account for the finding that the reported correlations were mainly present for participants better fit by the SWIM? Here, another finding might provide an answer. Across both studies in Ashby and Rakow (2014), participants better fit by the SWIM drew substantially more samples than did those better fit by the VUM (see Fig. 2). Now note that the impact of noise on the amount of underestimation increases with sample size; specifically, the lines in Figure 2 are not linear, but deviate more from the actual window size as the actual sample size increases. To illustrate, with an actual sample size of 6 (the average sample size of the people best fit by the VUM), and a noise level of 0.8, the underestimation is 15% (5.5 out of 6), whereas with an actual sample size of 16 (the average sample size for people best by the SWIM), the underestimation is 20% (12.8 out of 16). As a consequence, for a given range of noise levels, the variability in estimated window sizes will be larger for large sample sizes than for small sample sizes—and thus also for participants better fit by the SWIM than for those better fit by the VUM. Consistent with this possibility, the variance of the window size estimates for the participants in Study 2 who were better fit by the SWIM (according to the original results) was larger than the variance for participants who were better fit by the VUM, with average standard deviations of 18.9 and 4.4, respectively,  $F(15, 35) = 7.62, p < .001$ . The smaller variance for the participants classified as VUM users may thus have concealed the pairwise relationships between window size, working memory, and response times that were found for participants identified as SWIM users. That the relationship between working memory capacity and estimated window size was present in the latter, but absent in the former, does thus not necessarily mean that the two groups of participants relied on different processes.

In summary, the potential confound of window size estimated by the SWIM and judgment noise along with the finding that model performance depends considerably on sample size challenge Ashby and Rakow's (2014) conclusion in favor of an all-or-nothing, end-of-sequence process in valuations from experience.

### **Conceptual Issues with End-of-Sequence Processing in Self-Terminated Search**

Beyond the statistical issues addressed so far, it is also instructive to assess the notion of an end-of-sequence evaluation process in valuation from experience on a conceptual level. In the sampling paradigm used by Ashby and Rakow (2014), participants sequentially draw samples from an initially unknown payoff distribution, and it is up to them to decide when to stop sampling. Should one expect decision makers to construct a valuation only after sampling has been terminated—as predicted by a strict interpretation of an end-of-sequence process? Note that, according to this approach, information processing takes place only after all information has been collected. This implies that a decision about the number of samples drawn should be unrelated to the sampled outcomes. Based on the evidence presented by Ashby and Rakow, however, this does not seem very plausible. In particular, sample size was found to be correlated with the variance of the lotteries (for similar findings, see Lejarraga, Hertwig, & Gonzalez, 2012; Pachur & Scheibehenne, 2012): in lotteries with a larger variance, participants drew a larger number of samples. Clearly, in order for sampling effort to be sensitive to the characteristics of individual lotteries (e.g., their variances), some form of on-line processing has to occur.

Another reason that speaks against end-of-sequence processing is Ashby and Rakow's (2014) finding that the data in the sampling paradigm display a recency effect (i.e., that more weight is given to more recent experiences; see also Pachur & Scheibehenne, 2012; Wulff, Hills, & Hertwig, 2014). Reviewing studies with paradigms that encourage end-of-sequence

processing (i.e., where people were presented with a sequence of evidence and asked for an evaluation at the end of the sequence), Hogarth and Einhorn (1992) reported that 34 of 54 studies (63%) showed a *primacy* effect, not a recency effect. Recency effects, by contrast, were found predominantly in studies explicitly enforcing step-by-step processing (in 20 of 22 studies; 91%). Based on these findings in the literature on valuations from experience, if Ashby and Rakow's participants had indeed relied on an end-of-sequence process, one would have expected a primacy effect to occur.

Given that people are not explicitly instructed to conduct step-by-step processing in the sampling paradigm, one may ask what leads to reliance on such a process. One interesting possibility is that a step-by-step process is triggered by the requirement in the sampling paradigm for the decision maker to decide when to stop sampling (note that in typical end-of-sequence studies, participants are presented with a sequence of outcomes of fixed length—that is, they do not have to decide when to terminate search; Hogarth & Einhorn, 1992). If this were the case, end-of-sequence processing should be more dominant in a modified sampling paradigm in which participants are presented with a sequence of outcomes of fixed length.

### **Interim Summary**

First, a more refined approach to the model evaluation conducted by Ashby and Rakow (2014) provided support for the SUM, rather than the SWIM. Second, under realistic levels of noise, the window size estimated by the SWIM substantially undershot the number of samples drawn, casting doubt on the conclusion that small window sizes reflect forgetting or constraints of working memory capacity. Third, the notion of a strict end-of-sequence process is inconsistent both with the finding of variance-sensitive sampling in the present and previous investigations (e.g., Pachur & Scheibehenne, 2012) and with a large body of research on belief updating (Hogarth & Einhorn, 1992).

Despite a reasonably large sample containing 96 participants and numerous data points per participant, suggesting considerable power to identify the underlying mechanisms, none of the tested models received clear support. We next consider two candidate explanations for this result.

## **What Underlies the Results of Ashby and Rakow (2014)?**

### **Individual Differences in the Valuation Process**

Our reanalysis of Ashby and Rakow's (2014) model evaluation showed that most people were best fit by the SUM, but that the aggregate AIC<sub>c</sub> and BIC values essentially showed no differences between the models. This finding suggests that some people may in fact have relied on processes akin to the VUM and the SWIM, while the majority relied on the SUM. One way to assess whether there is evidence for such a split in processes is to evaluate the evidence for the three models not on the aggregate level, but on the individual level. Another way to test whether classifications are valid is to inspect the sampling and valuation behavior of individuals with respect to their model classification. The previous analyses have highlighted that (a) noise levels were generally high and (b) model classification may hinge on the number of samples taken (see Fig. 2). Because none of the models is theoretically linked to high noise or particular sample sizes, findings indicating that those classified confidently exhibit peculiar behavior in terms of extreme noise or sample size may thus further qualify the classification.

First, we assess the evidence based on BIC and AIC<sub>c</sub> values on the level of the individual. To this end, we examine the evidence for the three models (i.e., SUM, VUM, and SWIM) using model weights. Model weights are defined as

$$w_M = \frac{e^{-\frac{1}{2}\Delta_{crit}M}}{\sum_i e^{-\frac{1}{2}\Delta_{crit}t_i}}, \quad (3)$$

where  $\Delta crit$  is the difference of model  $M$  to the best performing model (among the set of competing models) on the respective information criterion (i.e., AIC<sub>c</sub> or BIC; see Lewandowsky & Farrell, 2010). Model weights vary between 0 and 1, with values of  $1/N_{\text{models}}$  and 1 indicating chance and perfect performance, respectively. Model weights can be evaluated by evidence categories proposed for Bayes factors (Jeffreys, 1961; Kass & Raftery, 1995; Wagenmakers, 2007), as both are directly related:

$$w_M = \frac{1}{1 + BF_{NM}} = \frac{1}{1 + e^{-\frac{1}{2}(\text{crit}_N - \text{crit}_M)}}, \quad (4)$$

where  $BF_{MN}$  denotes the Bayes factor for model  $M$  over a competing model  $N$ . Table 2 shows the evidence categories for the three competing models (adapted from Kass & Raftery, 1995). In extending the evidence categories to the comparison of three models, some rescaling was necessary to ensure that a Bayes factor of 1 equates to a model weight of .33, which denotes chance level.<sup>9,10</sup>

Table 2. *Evidence Categories for Bayes Factors and Model Weights ( $w_m$ ) for the Comparison of Three Models. Adapted from Kass and Raftery (1995, see also Wagenmakers, 2007).*

---

<sup>9</sup> Because Bayes factors express the relative evidence for the comparison of only two models and not three as for our model weights, a simplifying assumption needed to be introduced to link the two together. We chose to assume identical  $\Delta crit$  for the second and third best models. Thus, we linked a particular Bayes factor, say  $BF = 3$ , to the model weight that would be obtained if both the second and third placed models deviated from the best model by exactly the  $\Delta crit$  that would lead to a Bayes factor of 3 in pairwise comparisons.

<sup>10</sup> Although these evidence categories originate from a Bayesian framework of model evaluation, we apply them to BIC and AIC<sub>c</sub> alike. This is justified because model selection based on AIC<sub>c</sub>, despite its information-theoretic origins, can be rendered perfectly consistent with a Bayesian analysis by assuming a *savvy* prior that reflects the number of parameters and data points (Burnham & Anderson, 2004).

Bayes factor	$w_M$	Evidence
1–3	.33–.6	Weak
3–20	.6–.83	Moderate
> 20	.83–1	Strong

Figure 3 displays the model weights for each participant in both studies by Ashby and Rakow (2014), ordered by and highlighting the highest model weight for a given participant. As can be seen, model weights for the SWIM and VUMr vary considerably across participants, showing strong evidence for some participants and very weak evidence for others. Importantly, the model weights are at least moderate for the SWIM for only eight participants and for the VUMr for three participants, allowing an unequivocal classification of these participants to the respective models (for both BIC and AIC<sub>c</sub>). Aside from these eleven participants, unequivocal classifications to either VUM and SWIM are not possible (given the level of evidence). Thus, participant-level model weights only appear to provide evidence for the use of the SWIM and VUMr in a few of the cases.



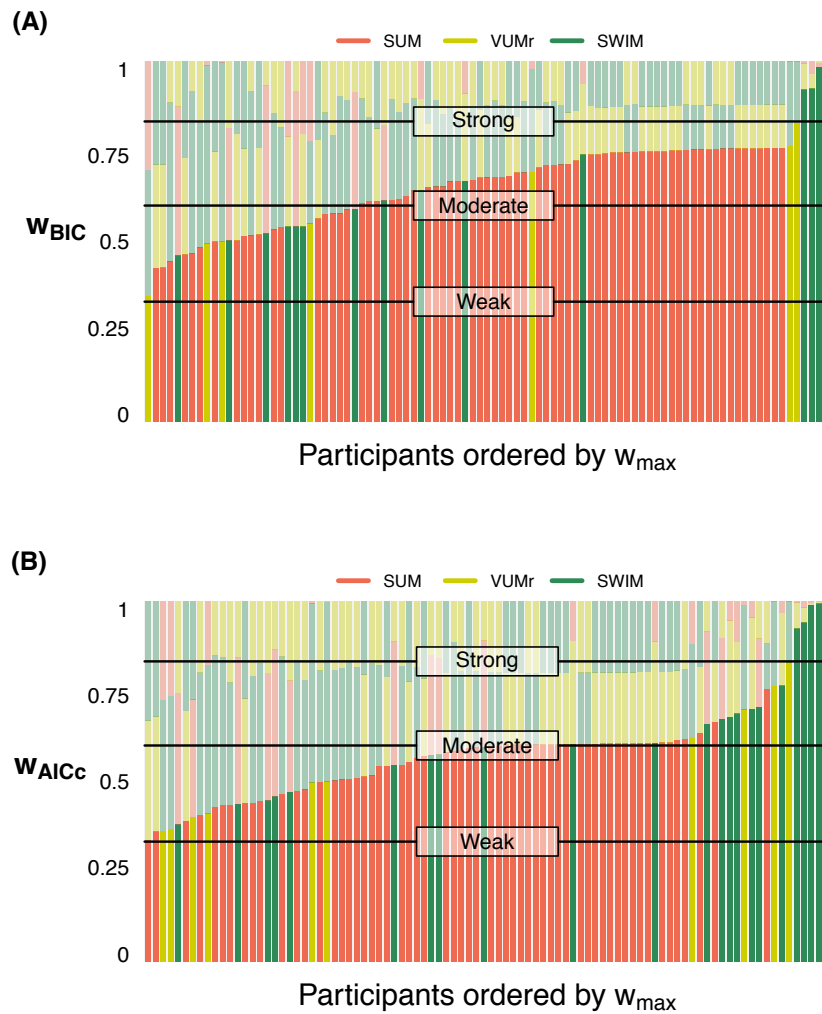


Figure 3. Model weights for each individual participant in Ashby and Rakow (2014). Panel (A) and panel (B) show model weights calculated for BIC and  $AIC_c$ , respectively. Levels of evidence were adapted from Kass and Raftery (1995; see Table 2).

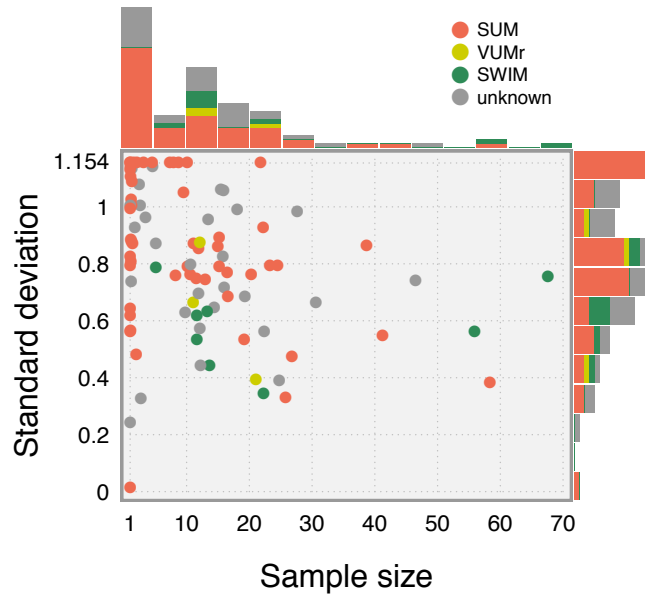
For the SUM, on the other hand, many model weights indicate good evidence. Note that because the SUM is nested within the SWIM and VUMr, there is an upper bound on

model weights in favor of the SUM (.76 for BIC and .6 for AIC<sub>c</sub>).<sup>11</sup> Thus, although only few participants could be confidently assigned to the VUMr or the SWIM, the proportion of participants best fit by the SUM reaches levels close to the theoretical maximum. The analysis of individual model weights thus provides some evidence that participants may have relied on different strategies.

Now we turn to the second analysis to assess whether some of those classifications have potentially been confounded with the behavior exhibited by the participants. To this end, we plot in Figure 4 the observed mean sample sizes against the estimated standard deviation of the judgment error distribution (based on the respective best fitting models); participants for whom the evidence for a respective model is at least moderate in respect to BIC are highlighted. The figure shows that the majority of participants identified as SUM users either sampled very little, or exhibited very high levels of judgment noise, or both. Under such conditions, it is questionable whether the SUM classification is valid, as it is unlikely that recency can take effect in so few samples. That is, a potentially true VUMr or SWIM with recency may not have been able to influence data such that it could be recovered. Importantly, 11 participants sampled only once at every lottery, rendering the superiority of the SUM a statistical necessity in these cases (as recency was impossible). Similar reservations hold for very high levels of judgment noise. For a different set of 11 participants, the estimated noise level was so high (larger than 1.15) that basically all possible valuations (ranging between 0 and 4) were equally likely. In these cases, the SUM necessarily wins because all models fit equally poorly and the SUM is punished less severely for its number of free parameters.

---

<sup>11</sup> The fact that the SUM is nested within the VUMr and SWIM allows the VUMr and SWIM to fit the data equally well. This naturally reduces possible differences on AIC<sub>c</sub> and BIC to the value the respective criteria assigned to having one additional parameter (i.e., about 2.22 for the AIC<sub>c</sub> and 3.69 for the BIC, based on 40 data points). The model weights resulting from these values, .6 for AIC<sub>c</sub> and .76 for BIC, place upper bounds for all model weights in favor of the SUM.



*Figure 4.* Sample size and judgment noise measured as the estimated standard deviation under the best fitting model for the participants in both studies by Ashby and Rakow (2014). Colors show model classifications based on at least moderate evidence provided by model weights calculated for BIC.

In sum, many participants who seriously explored the lotteries (as indicated by sample sizes larger than 1) and whose valuations exhibited some correspondence to the experienced mean (as indicated by standard deviations smaller than 1.1) could not be classified as relying on one of the three tested valuation processes (although we used the rather lenient criterion of at least moderate evidence in BIC-based model weights). Thus, the analysis of individual model weights supported the hypothesis that participants may have relied on different processes. However, it also failed to provide clear evidence for one of the models in terms of the majority of participants behaving “reasonably.” We next examine the possibility that the

discrimination between models is generally hampered due to lack of information in the data and complexity differences in the models.

### **Model Complexity**

The ability to discriminate between models and thus the ability to identify the underlying process depends on the design, the data created by it, the set of models under consideration, and the measures used to assess their performance. To enable discrimination between models, a design must generate data such that the models often make different predictions. For valuations from experience using the sampling paradigm, this is difficult to achieve, as people's active search can lead to uninformative data when search is terminated before more than two different outcomes have been observed. If search is stopped after only one outcome has been sampled, then no matter how models are compared, it is impossible to discriminate between them. To illustrate, with the two-outcome lotteries used by Ashby and Rakow (2014) and the observed distribution of mean sample sizes over both studies, this occurred in 34% of trials (matching its expectation of 32% as determined by simulation). More importantly, for participants who sampled no more than twice (on average), it is expected to happen in almost 89% of trials. Hence, for these participants, only 4 of the 40 lotteries can be used to potentially discriminate between the models. But even if two or more different outcomes are observed, the sequence of samples can still make it impossible to discriminate between models—for instance, when outcomes are relatively evenly distributed.<sup>12</sup>

---

<sup>12</sup> When outcomes are distributed relatively evenly across the sequence, different degrees of recency, no recency, or primacy will lead to very similar valuations. For illustration, the sequence {1, 2, 1, 2, 1, 2, 1} will lead to highly similar predictions irrespective of whether, for instance, the valuation is based on the three most recent outcomes alone (as in the SWIM), or on a gradual weighting over all outcomes (as in the VUMr). In the context of the SWIM, it is, analogously, very difficult to discern which subset of the sequence underlies the valuation.

Model discrimination can also fail for reasons lying within the formulation of the models. For the present analysis at least two issues deserve mentioning. First, all three models predict the expected value of the lottery. Due to the law of large numbers, the models will thus make increasingly similar predictions with growing sample sizes.<sup>13</sup> Because models making identical predictions can be distinguished only on the basis of the number of parameters, the SUM is increasingly likely to emerge as the winner as sample size goes up. Second, the complexity of the VUMr and the SWIM clearly depends on the sample size. To illustrate, in the simple case of just two observed values, e.g.,  $\{1, 3\}$ , the VUMr can perfectly fit any valuation between 2 and 3 by shifting the  $\phi$ -parameter between 0 and 1. The SWIM, on the other hand, can only make exactly two predictions with two samples, i.e., 2 ( $\zeta = 2$ ) and 3 ( $\zeta = 1$ ). For very small sample sizes, the VUMr can thus be expected to perform much better in fitting empirical data than the SWIM. For large sample sizes, however, the reverse may be true. The VUMr is required to use all samples for any  $\phi$  larger than 0, which may place a large restriction on the range of valuations it can fit. Due to its ability to ignore entire subsets of the data, the SWIM may prove to be more flexible here.

To examine the extent to which each of the models can nonetheless be correctly identified, we conducted a model recovery analysis based on the setup of Ashby and Rakow (2014). Specifically, we simulated 1,000 agents for every combination of 20 levels of sample size and 20 levels of judgment noise (size =  $[2, 40]$ , noise =  $[\cdot 2, 1.1]$ , both covering 95% of

---

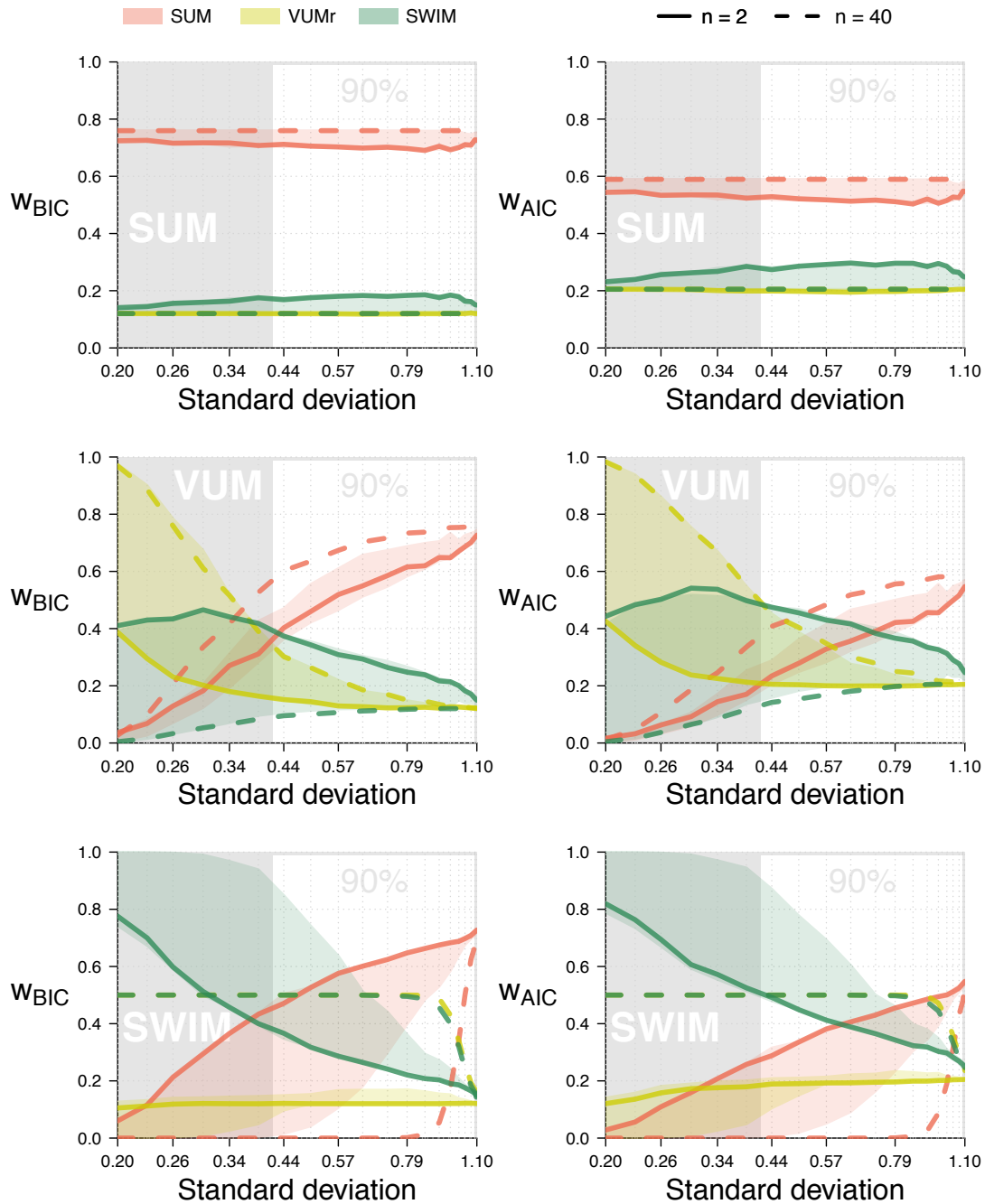
<sup>13</sup> For the VUMr and SUM, this holds immediately. For any given recency parameter that is larger than 0, the prediction of the VUMr will approach the prediction of the SUM with increasing sample size. For the SWIM, the same holds when recency in the SWIM is conceptualized to be proportional to the total sample size. Although not explicitly formulated in that way, the high positive correlation between window size and sample size reported by Ashby and Rakow (2014) suggests such a proportional relationship. The convergence in the limit implies that recency mechanisms are less likely to be detected for larger sample sizes, because under these circumstances the predictions of the simpler SUM will lie closer to those of the VUMr and SWIM.

the observed values) for each of the models as the generating process. For the VUM and the SWIM, we assumed levels of recency that match the median parameter estimates obtained from the Ashby and Rakow data (VUMr:  $\phi = .77$ ; SWIM:  $\zeta/n = .73$ ). Given the empirical finding that window and sample sizes are strongly correlated, for the SWIM the window size producing the valuation was determined as a fraction of the agent's sample size. We then fit the SUM, VUMr, and SWIM to each agent's data, and recorded their relative performance in terms of model weights based on AIC<sub>c</sub> and BIC.

Figure 5 shows median model weights for the three models when the SUM (upper panel), the VUMr (middle panel), or the SWIM (lower panel) was the generating mechanism, separately for the different levels of noise (shown on the x-axis). The line types and transparent shapes illustrate the effect of varying sample sizes. The dashed and solid lines represent mean sample sizes of 2 and 40, respectively. The transparent shapes illustrate the range (minimum to maximum) of observed median model weights for mean sample sizes between 4 and 38. The white background highlights the range that covers 90% of the observed noise levels for the participants in both studies by Ashby and Rakow (2014).

As evidenced by many misclassifications, the results in Figure 5 highlight that using the Ashby and Rakow (2014) design to identify the underlying model for valuations from experience can be a rather thorny endeavor. Specifically, model recovery not only seems error-prone, but its accuracy seems to differ considerably across the models and to depend on the behavior of the agents. First, for the vast majority of conditions implemented in our simulation, the SWIM appears to be more flexible than the VUMr, particularly when the sample size of observations is large. As a result, valuations generated by the VUMr are more likely to be incorrectly attributed to the SWIM than vice versa. The only exceptions occurred with a sample size of 2, due to the fact that the VUMr can perfectly mimic the two possible

predictions of the SWIM for a sample size of 2 (by assuming  $\phi = 1$  or  $\phi = 0$ ). Second, with that same exception, the true model is generally much better identified when sample size is small. This implies that nondiagnostic data, which are more likely to occur for small samples, present a smaller problem than the issues of sample size-dependent flexibility and converging predictions between the SUM, VUMr, and SWIM for increasing sample sizes. Importantly, this further implies that the observed inaccuracies cannot be easily solved by design improvements—for instance, by increasing the number of problems or forcing people to sample more. Instead, this finding calls for methods that properly account for the models' actual flexibility, an issue to which we return later.



*Figure 5.* Model recovery as a function of criterion, judgment noise (standard deviation), and sample size. Panels in the upper, middle, and lower rows show the fits for data generated by the SUM, VUMr, and SWIM, respectively. Panels on the left show fits expressed as median model weights based on BIC; panels on the right, based on AIC<sub>c</sub>. The solid lines indicate the performance for a sample size of 2; the dashed lines, for a sample size of 40. The transparent shapes show the range of performance for mean sample sizes between 4 and 38.



In sum, based on the results of the model recovery, we can conclude that it was not SUM alone that generated the data in Ashby and Rakow's (2014) studies—otherwise, as can be seen from the recovery analysis for the SUM, it would have been the clearly dominating model for every participant. However, it is nearly impossible to discern whether few, some, or all participants actually relied on the SUM, the VUM or the SWIM. Thus, any conclusion regarding whether the SUM, the VUM, or the SWIM is the best model that is drawn on the basis of Ashby and Rakow's study design and standard indices such as BIC and AIC/AIC<sub>c</sub> must be treated with great caution.

### **Challenges for Studying and Modeling Valuations from Experience**

#### **Reducing Noise, Improving Fit**

Our analyses highlighted that it was often virtually impossible to infer whether participants relied on the SUM, VUM, or SWIM in the Ashby and Rakow (2014) data. The foremost reason was very high levels of noise. Future studies should thus aim to identify the source of noise and attempt to reduce it.

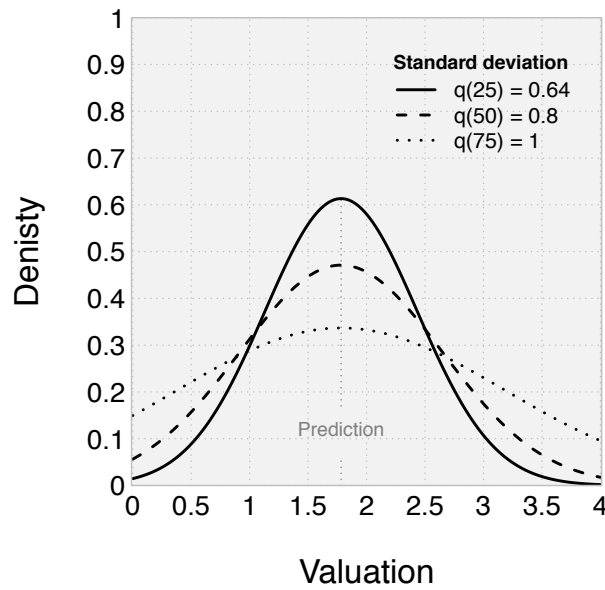
One way to achieve this would be to make sure that participants completely understand the Becker-DeGroot-Marschak procedure (Becker, DeGroot, & Marschak, 1964), an issue that has been discussed in the pertinent literature (Irwin et al., 1998; Krahnen, Rieck, & Theissen, 1997; James, 2007; Safra, Segal, & Spivak, 1990). Given that noise reflects unsystematic behavior, another avenue would be to test the use of other noise distributions (e.g., a beta distribution or Student's *t* distribution) that can handle extreme valuations more

easily.<sup>14</sup> Moreover, judgment noise has also sometimes been found to depend on the lottery (e.g., Stott, 2006); the precision of a model might thus be improved by implementing noise such that it is sensitive to the characteristics of a lottery (e.g., its variance), rather than assuming a constant error.

However, the sheer magnitude of noise found for the Ashby and Rakow (2014) data suggests the existence of factors beyond unsystematic behavior (i.e., a trembling hand). To illustrate this, we plot in Figure 6 the predicted distribution of valuations for the SUM under noise levels matching the first, second, and third quartile found in the data. Clearly, with such noise levels, many valuations are rendered consistent with the model. This implies in turn that many marked deviations from the models' predictions had to be accommodated, which led to increased standard deviations. In fact, the mean absolute deviations for the best fitting models across all valuations amounted to .73 for the SUM, .67 for the VUM, and .66 for SWIM.

---

<sup>14</sup> We tested a beta, truncated  $t$  distribution in two parameterizations ( $t1$ :  $\mu$  and  $\nu$ ;  $t2$ :  $\mu$  and  $\sigma$  with  $\nu = 1$ ) and found that the beta based on maximum likelihood outperformed the  $t$  and normal distributions for the data of Ashby and Rakow (2014).



*Figure 6.* Illustration of judgment noise for noise levels matching the first, second, and third quartile of the noise levels found for the Ashby and Rakow (2014) data for a mean prediction of 1.78, which is the expectation across all lotteries.

Future studies should thus attempt to reduce noise by improving the models. In so doing, it is important to consider that both subjective expected utility and memory-based recency have found various formalizations in recent literature. For the sake of simplicity and to match Ashby and Rakow's (2014) investigation, we have not parameterized weighting of probability and outcome. Clearly, including parameters for outcome and probability transformations may help increase the fit of the models to the data (Fox & Poldrack, 2009; Tversky & Kahneman, 1992). Similarly, we have not attempted to implement other forms of recency. Particularly promising would be the construction of a valuation of experience model based on ACT-R principles (Anderson & Lebiere, 1998; see, e.g., Erev et al., 2010; Gonzales & Dutt, 2011, for applications to decisions from experience). In contrast to the models

implemented here, ACT-R offers a process informed by memory research, in which the retention and hence the utilization of a sample is a function not only of its absolute position, but also of its relative position to other sampled outcomes.

Another potentially relevant set of strategies may be given by the three participants excluded at the outset of our reanalysis of the Ashby and Rakow data (2014). Those three participants reported one of the observed values in almost all trials, despite sample sizes larger than one. Future studies should also consider that the reliance on the SUM, VUM, and SWIM or any other strategy may change from trial to trial.

### **Active Sampling**

Yet, for at least three reasons, the most important aspect in understanding valuations from experience is the participant's active role in sampling. First, by requiring a person to decide when to stop sampling, active sampling requires processing of the sampled information during sampling, as evidenced by the correlation of variance and sample size (Ashby & Rakow, 2014; Pachur & Scheibehenne, 2012). Such on-line processing is probably not independent of the processes recruited to form a final valuation. Rather it can be assumed to either inform or replace final end-of-sequence processes.

Second, if people are free to sample as much or as little as they want, leading to individual variability as well as variability across trials in search, it can be very difficult to correctly identify the underlying processes in valuations from experience. As shown for the VUM and the SWIM, the amount of sampling can affect the relative complexity of models (Grünwald, Myung, & Pitt, 2005). As a consequence, one model can outperform the other for small sample sizes, but be outperformed for large sample sizes, relatively irrespective of which of the two is the true underlying process.

Finally, active sampling hampers the recovery of the underlying processes in valuations from experience by affecting the diagnosticity of the data. If a person samples only a few times, the chances are high that only one of the lotteries' outcomes is observed, which renders predictions of the models identical and their discrimination impossible. Problems also arise if a person samples too many times. Models such as the SUM, VUM, and SWIM with identical expectations will converge in their prediction when sample size gets large, resulting in the same problem as for model discrimination. In addition, the stochastic aspect of sampling can render sequences of any length more or less informative for a given set of models. For instance, sequences in which outcomes are relatively evenly or symmetrically distributed are consistent with recency and primacy of the same degree, and will likely fail to result in different predictions for all-or-nothing versus gradual forms of recency, as implemented by the SWIM and VUM, respectively.

In sum, active sampling strongly influences the psychological processes involved, as well as the methods applied to uncover them. The statistical issues can largely be addressed by more advanced designs and methods. We highlight potential measures in the next section. The impact of active sampling on the cognitive process triggered, however, needs to be considered theoretically. A useful starting point is given by recent research on self-directed learning (Gureckis & Markant, 2012; Markant & Gureckis, 2014). This research aims to explain why and how active information search (as opposed to passive observation) can lead to different and often better performance (e.g., in category learning). Recent attempts to model the process of hypothesis generation during the processing of fixed sequences of information may be equally relevant (Jahn & Braatz, 2014; Lange, Thomas, & Davelaar, 2012; Thomas, Dougherty, Sprenger, & Harbison, 2008).

### **Improving Model Discriminability**

We see three avenues for improving experimental designs testing models of valuations from experience. First, a larger number of lotteries could be used, as this will increase the chance of observing diagnostic sequences. Second, participants could be motivated to sample more (to increase the chance of observing more than one outcome or reducing convergence in predictions), for instance, by means of incentives (see Hau, Pleskac, Kiefer, & Hertwig, 2008). Third, instead of using the common two-outcome lottery, researchers could use multi-outcome or even continuous lotteries (Wulff et al., 2014). This will not only reduce the occurrence of sequences in which the same outcome is drawn multiple times, but also increase the richness of sequences of all lengths. Naturally, all three measures may change participants' behavior in complex ways.

To address the issue of model discriminability, future investigations should consider applying more advanced model comparison methods, such as Bayes Factors (BF; Kass & Raftery, 1995; see Scheibehenne, Rieskamp, & Wagenmakers, 2013, for practical applications) or Normalized Maximum Likelihood (NML; Myung, Navarro, & Pitt, 2005; see Kellen, Klauer, & Bröder, 2013). These methods are better able than comparisons based on BIC and AIC to take model complexity into account. In fact, BIC and AIC are constrained solutions of the Bayesian and information theoretical frameworks, respectively. Admittedly, applying BF and NML can be difficult, as computing the marginal likelihood (for BF) requires integration over the parameter space, and determining model complexity (for NML; Grünwald et al., 2005) requires integration over the data space. Approximate methods have been developed (Gamerman & Lopes, 2006; Grünwald et al., 2005; Kass & Raftery, 1995).

Complementing these formal approaches, Monte Carlo methods similar to the approach taken in this paper have been developed to shed light on model mimicry, recovery, and optimal study design (e.g., Navarro, Pitt, & Myung, 2004; Wagenmakers, Ratcliff,

Gomez, & Iverson, 2004). One particularly intuitive and easily implemented method is landscaping (Navarro et al., 2004). Similar to a model recovery approach, landscaping addresses the issue of model complexity by evaluating a model's performance on real data relative to its performance on data generated by multiple models under many different parameter combinations. Importantly, Navarro and colleagues (2004) have also demonstrated how this method can be applied to assess the discriminability of design variants. Thus, future studies could assess the impact of our suggestions for design improvement without collecting actual data, while at the same time properly controlling for the models' true complexity.

### **Conclusion**

With the SWIM, Ashby and Rakow (2014) proposed an interesting alternative to the traditional approach to model recency in valuations from experience, and its assumption that more distance experiences have a gradually decreasing influence on the valuation of an object. We welcome their proposal as it highlights the overdue need to develop and rigorously compare models of experience-based judgment and decision making. To this end, our analysis highlights a number of challenges in current approaches to testing and implementing models of valuations from experience. We suggest several improvements to address these challenges, some of which will admittedly not be easy to implement in every situation. We nevertheless hope that future research will take up the challenge, as active information search prior to valuation or choice is a ubiquitous real world scenario.

### References

- Ackerman, P. L., Beier, M. E., & Boyle, M. O. (2005). Working memory and intelligence: The same or different constructs? *Psychological Bulletin*, 131, 30–60.
- Akaike, H. (1973). Information theory as an extension of the maximum likelihood principle. In B. N. Petrov & F. Csaki (Eds.), *Second international symposium on information theory* (p. 267-281). Budapest: Akademiai Kiado.
- Anderson, J. R., & Lebiere, C. (1998). *The atomic components of thought*. Mahwah, NJ: Erlbaum.
- Ashby, N. J., & Rakow, T. (2014). Forgetting the past: Individual differences in recency in subjective valuations from experience. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 40, 1153–1162.
- Baddeley, A. (2012). Working memory: Theories, models, and controversies. *Annual Review of Psychology*, 63, 1–29.
- Becker, G. M., DeGroot, M. H., & Marschak, J. (1964). Measuring utility by a single-response sequential method. *Behavioral Science*, 9, 226–232.
- Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach*. New York: Springer.
- Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods and Research*, 33, 261–304.
- Bush, R., & Mosteller, F. (1955). *Stochastic models for learning*. New York, NY: Wiley.
- Cowan, N. (2000). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, 24, 87–185.
- Ebbinghaus, H. (1885/1913). *Memory: A contribution to experimental psychology*. New York, NY: Teachers College Press. (Original work published 1885)



- Erev, I., Ert, E., Roth, A. E., Haruvy, E., Herzog, S. M., Hau, R., Hertwig, R., Stewart, T., West, R., & Lebiere, C. (2010). A choice prediction competition: Choices from experience and from description. *Journal of Behavioral Decision Making*, 23(1), 15–47.
- Fox, P. (1997). The Port Mathematical Subroutine Library, Version 3. Murray Hill, NJ: AT&T Bell Laboratories. Retrieved from: <http://www.bell-labs.com/project/PORT/>.
- Fox, C. R., & Poldrack, R. A. (2009). Prospect theory and the brain. In P. W. Glimcher, E. Fehr, C. Camerer, & R. A. Poldrack (Eds.), *Neuroeconomics: Decision making and the brain* (pp. 145-174). London, UK: Elsevier.
- Gamerman, D., & Lopes, H. F. (2006). *Markov chain Monte Carlo: stochastic simulation for Bayesian inference*. CRC Press.
- Gonzalez, C., & Dutt, V. (2011). Instance-based learning: Integrating sampling and repeated decisions from experience. *Psychological Review*, 118(4), 523–551.
- Grünwald, P. D., Myung, I. J., & Pitt, M. A. (Eds.) (2005). *Advances in minimum description length: Theory and applications*. Cambridge, MA: MIT Press.
- Gureckis, T. M., & Markant, D. B. (2012). Self-directed learning: A cognitive and computational perspective. *Perspectives on Psychological Science*, 7, 464–481.
- Hastie, R., & Park, B. (1986). The relationship between memory and judgment depends on whether the judgment task is memory-based or on-line. *Psychological Review*, 93, 258–268. doi:10.1037/0033-295X.93.3.258
- Hau, R., Pleskac, T. J., Kiefer, J., & Hertwig, R. (2008). The description–experience gap in risky choice: The role of sample size and experienced probabilities. *Journal of Behavioral Decision Making*, 21, 493–518.
- Hertwig, R. (in press). Decisions from experience. In G. Keren & G. Wu (Eds.), *Blackwell handbook of decision making*. Oxford, UK: Blackwell.
- Hertwig, R., Barron, G., Weber,

- E. U., & Erev, I. (2006). The role of information sampling in risky choice. In K. Fiedler & P. Juslin (Eds.), *Information sampling and adaptive cognition* (pp. 75–91). New York, NY: Cambridge University Press.
- Hertwig, R., & Erev, I. (2009). The description–experience gap in risky choice. *Trends in Cognitive Sciences*, 13(12), 517–523.
- Hills, T. T., & Hertwig, R. (2010). Information search in decisions from experience: Do our patterns of sampling foreshadow our decisions? *Psychological Science*, 21(12), 1787–1792.
- Hogarth, R. M., & Einhorn, H. J. (1992). Order effects in belief updating: The belief-adjustment model. *Cognitive Psychology*, 24(1), 1–55.
- Irwin, J. R., McClelland, G., McKee, M., Schulze, W. D., & Norden, N. E. (1998). Payoff dominance vs. cognitive transparency in decision making. *Economic Inquiry*, 36(2), 272–285.
- Jahn, G., & Braatz, J. (2014). Memory indexing of sequential symptom processing in diagnostic reasoning. *Cognitive Psychology*, 68, 59–97.
- James, D. (2007). Stability of risk preference parameter estimates within the Becker-DeGroot-Marschak procedure. *Experimental Economics*, 10(2), 123–141.
- Jeffreys, H. (1961). *The theory of probability*. Oxford: Oxford University Press.
- Kane, M. J., Brown, L. H., McVay, J. C., Silvia, P. J., Myin- Germeys, I., & Kwapil, T. R. (2007). For whom the mind wanders, and when: An experience-sampling study of working memory and executive control in daily life. *Psychological Science*, 18, 614–621.
- Kane, M. J., & Engle, R. W. (2003). Working-memory capacity and the control of attention: The contributions of goal neglect, response competition, and task set to Stroop

- interference. *Journal of Experimental Psychology: General*, 132, 47–70.
- Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90, 773–795.
- Kellen, D., Klauer, K. C., & Bröder, A. (2013). Recognition memory models and binary-response ROCs: A comparison by minimum description length. *Psychonomic Bulletin and Review*, 20, 693–719.
- Krahnke, J. P., Rieck, C., & Theissen, E. (1997). Inferring risk attitudes from certainty equivalents: Some lessons from an experimental study. *Journal of Economic Psychology*, 18(5), 469–486.
- Lange, N. D., Thomas, R. P., & Davelaar, E. J. (2012). Temporal dynamics of hypothesis generation: The influences of data serial order, data consistency, and elicitation timing. *Frontiers in Psychology*, 3, 215.
- Lejarraga, T., Hertwig, R., & Gonzalez, C. (2012). How choice ecology influences search in decisions from experience. *Cognition*, 124, 334–342.
- Lewandowsky, S., & Farrell, S. (2010). *Computational modeling in cognition: Principles and practice*. London, UK: Sage.
- March, J. G. (1996). Learning to be risk averse. *Psychological Review*, 103, 309–319.
- Markant, D. B., & Gureckis, T. M. (2014). Is it better to select or to receive? Learning via active and passive hypothesis testing. *Journal of Experimental Psychology: General*, 143(1), 94–122.
- Myung, J. I., Navarro, D. J., & Pitt, M. A. (2006). Model selection by normalized maximum likelihood. *Journal of Mathematical Psychology*, 50, 167–179.
- Navarro, D. J., Pitt, M. A., & Myung, I. J. (2004). Assessing the distinguishability of models and the informativeness of data. *Cognitive Psychology*, 49(1), 47–84.

- Pachur, T., & Scheibehenne, B. (2012). Constructing preference from experience: The endowment effect reflected in external information search. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 38, 1108–1116.
- R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from <http://www.R-project.org/>
- Rubin, D. C., & Wenzel, A. E. (1996). One hundred years of forgetting: A quantitative description of retention. *Psychological Review*, 103, 734–760
- Safra, Z., Segal, U., & Spivak, A. (1990). The Becker-DeGroot-Marschak mechanism and nonexpected utility: A testable approach. *Journal of Risk and Uncertainty*, 3, 177–190.
- Scheibehenne, B., Rieskamp, J., & Wagenmakers, E. J. (2013). Testing adaptive toolbox models: A Bayesian hierarchical approach. *Psychological Review*, 120(1), 39–64.
- Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6, 461–464.
- Stott, H. P. (2006). Cumulative prospect theory’s functional menagerie. *Journal of Risk and Uncertainty*, 32, 101–130.
- Thomas, R., Dougherty, M., Sprenger, M., & Harbison, J. (2008). Diagnostic hypothesis generation and human judgment. *Psychological Review*, 115, 155–185.
- Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, 5, 297–323.
- Vrieze, S. I. (2012). Model selection and psychological theory: A discussion of the differences between the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). *Psychological Methods*, 17, 228–243.
- Wagenmakers, E. J. (2007). A practical solution to the pervasive problems of *p* values. *Psychonomic Bulletin & Review*, 14, 779–804.

- Wagenmakers, E. J., Ratcliff, R., Gomez, P., & Iverson, G. J. (2004). Assessing model mimicry using the parametric bootstrap. *Journal of Mathematical Psychology*, 48, 28–50.
- Wulff, D. U., Hills, T. T., & Hertwig R. (2014) How short- and long-run aspirations impact search and choice in decisions from experience. *Submitted manuscript*.
- Wulff, D. U., Hills, T. T., & Hertwig, R. (2014). Online product reviews and the description–experience gap. *Journal of Behavioral Decision Making*. Advance online publication. doi:10.1002/bdm.1841
- Yechiam, E., & Busemeyer, J. R. (2006). The effect of foregone payoffs on underweighting small probability events. *Journal of Behavioral Decision Making*, 19(1), 1–16.

## Appendix A

### Monte Carlo Simulation and Model Recovery

Simulations were based on the following data generation and estimation procedures.

#### Data Generation

For each in a set of lotteries—in Ashby & Rakow (2014), the set was 40 lotteries—a simulated participant first drew  $n$  samples. A valuation for the current lottery was then determined based on the sampled outcomes according to the generating model. The valuation resulted from a draw from a normal distribution that was truncated to prevent valuations smaller than 0 or larger than 4, with a mean equaling the prediction of the generating model and a standard deviation that was a function of the respective noise level,  $v \sim N_{trunc}(v_{\text{model}}, \sigma_{\text{noise}})$ . To match the modeling in Ashby and Rakow, we assumed noise to be constant across the lotteries.

#### Parameter Estimation

The models were fitted to the valuation for each simulated participant by maximizing log-likelihood over all  $J$  lotteries using a normally distributed noise:

$$LL = \log \left( \frac{\prod_j^J \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{1}{2} \left( \frac{v_j - m_j}{\sigma} \right)^2}}{\Phi(4, m_j, \sigma) - \Phi(0, m_j, \sigma)} \right) \quad (\text{A1})$$

with  $m_j$  being the predicted valuation of the model,  $v_j$  being the data, and  $\Phi$  being the cumulative normal probability distribution. VUM and SWIM were implemented as defined in Equations 1 and 2. The predicted valuation of SUM at the  $n^{\text{th}}$  sample, based across all  $n$  samples, was defined as

$$v_n = \frac{1}{n} \sum_i^n x_i. \quad (\text{A2})$$

The VUM parameters  $\sigma$  and  $\theta$  (see Equation 1) were jointly estimated using R (R Core Team, 2013) via a quasi-Newton minimization procedure from the PORT library (Fox, 1997). The same approach was applied for the SUM. To estimate the SWIM parameters, we used a different approach, as stable estimates were obtained only when the starting points for the window size were close to the true window size. Instead, we performed an exhaustive search for the window size  $\theta$  and optimized the LL for each discrete value of  $\theta$  to find the optimal value of  $\sigma$ , again using PORT routines. In the rare cases (less than 1% of runs) in which multiple window sizes led to the same optimal fit, the smaller value was chosen as it represents a more conservative assumption. To avoid local minima, we repeated the fitting for several start values for  $\sigma$  and  $\theta$ .

How Short- and Long-Run Aspirations Impact Search and Choice  
in Decisions from Experience

Dirk U. Wulff

Max Planck Institute for Human Development, Berlin, Germany

University of Basel, Switzerland

Thomas T. Hills

University of Warwick, Coventry, UK

Ralph Hertwig

Max Planck Institute for Human Development, Berlin, Germany

Author Note

Dirk U. Wulff, Center for Adaptive Rationality, Max Planck Institute for Human Development, Berlin, Germany; Thomas T. Hills, Department of Psychology, University of Warwick, Coventry, UK; Ralph Hertwig, Center for Adaptive Rationality, Max Planck Institute for Human Development, Berlin, Germany. We are grateful to Susannah Goss for editing the manuscript and we thank the German Research Foundation and the Swiss National Science Foundation for grants to the third author (HE 2768/7-2; 100014-130397).

Correspondence concerning this article should be addressed to Dirk U. Wulff, Center for Adaptive Rationality (ARC), Max Planck Institute for Human Development, Lentzeallee 94, 14195 Berlin, Germany. Phone: +49 30 82406 475. E-mail: wulff@mpib-berlin.mpg.de



**Abstract**

To what extent do people adapt their information search policies and subsequent decisions to the long- and short-run consequences of choice environments? To address this question, we investigated exploration and exploitation policies in choice environments that involved single or multiple plays. We further compared behavior in these environments with behavior in the standard sampling paradigm. Frequently used in research on decision from experience, this paradigm does not explicitly implement the choice in terms of the short or long run. Results showed that people searched more in the multi-play environment than in the single-play environment. Moreover, the substantial search effort in the multi-play environment was conducive to choices consistent with expected value maximization, whereas the lesser search effort in the single-play environment was compatible with the goal of minimizing the risk of winning nothing. Furthermore, choice and search behaviors in the sampling paradigm predominantly echoed those observed in the single-play environment. This suggests that, when not instructed otherwise, participants in the sampling paradigm appear to favor search and choice strategies that embody short-run aspirations. Finally, the present findings challenge the revealed preference approach in decisions from experience, while also suggesting that information search may be an important and potentially even better signal of preference or aspirations than choice.

*Keywords:* decisions from experience, information sampling, risky choice, one-shot and multi-play gambles

## 1. Introduction

Choices between uncertain options can be interpreted as representing either one-shot or repeated decisions. A lottery ticket, for instance, represents a single-play decision; its entry price entitles the player to exactly one play of the lottery. A choice to buy car insurance, on the other hand, guarantees against repeated plays of a gamble that is realized each time the car is driven. More generally, decisions to buy products that will be consumed either once (e.g., a dinner in a gourmet restaurant) or many times (e.g., a pair of running shoes) involve different time horizons. These may, in turn, prompt differences in the decisions made as well as in the information needed to render a decision. For illustration, consider the offer that Nobel-prize winning economist Paul Samuelson (1963) once made his lunch partners: “to bet each \$200 to \$100 that the side of the coin *they* specified would not appear at the first toss” (p. 50). One colleague, whom Samuelson identified as a distinguished scholar but otherwise granted anonymity, responded to the offer by saying: “I won’t bet because I would feel the \$100 loss more than the \$200 gain. But I’ll take you on if you promise to let me make 100 such bets” (p. 2). Samuelson (1963) considered his colleague’s preference to be inconsistent with expected utility theory and, by extension, to be irrational (a fallacy of large numbers): “... no sequence is acceptable if each of its single plays is not acceptable” (p. 3).

More recent analyses, however, have concluded that models of expected utility theory—by many considered *the* normative theory of individual decision making—can in fact capture the colleague’s preference for safety in numbers, assuming that the 100 bets are aggregated to a single choice. Ex ante aggregation brings the final distribution of potential outcomes of a gamble much closer to its expected value and accordingly reduces the likelihood of losses (Aloysius, 2007; Kahneman & Lovallo, 1993; Wedell, 2011; see also Peköz, 2002). Thus, in decisions under uncertainty, single-play and multi-play choice

environments effectively entail different payoff distributions. More importantly, in decisions under risk where options' outcomes and probabilities are explicitly described (e.g., \$5000 with probability .09, otherwise \$0), people favor choices consistent with expected utility maximization in multi-play environments, but less so in single-play environments (Montgomery & Adelbratt, 1982; Wedell, 2011).

What is less well understood—and the focus of this article—is how people respond to single- and multi-play environments in which they first have to search for information before making a choice. We address this question by implementing the two choice protocols described in Samuelson's anecdote within the sampling paradigm, a popular design used in research on decisions from experience (Hertwig, Barron, Weber, & Erev, 2004; Hertwig & Erev, 2009). In the sampling paradigm, people first explore the possible outcomes of risky options in a self-directed and self-terminated sampling process before making a decision based on their sampled experience.

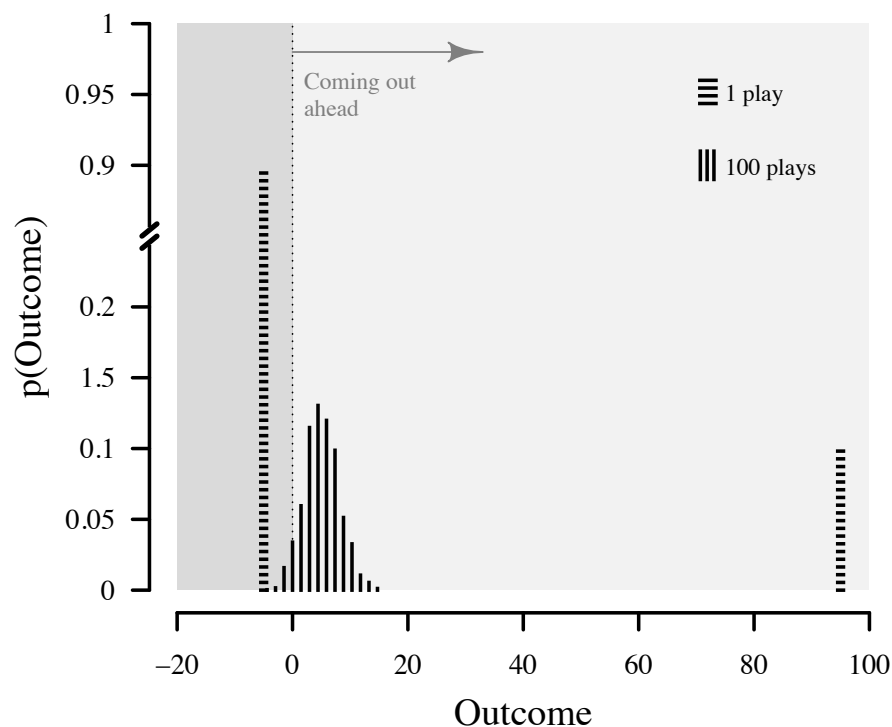
By clearly separating information search and choice, we can add search as a new dimension to the analysis of the effects of single-play and multi-play choice environments (DeKay & Kim, 2005; Montgomery & Adelbratt, 1982; Redelmeier & Tversky, 1992; see Wedell, 2011, for a brief review). In particular, information search can uncover differences in the motivational approach to single- and multi-play choice environments in the form of short- and long-run aspirations. Our investigation will also permit us to further analyze a recently discovered relationship between information search and choice that may originate from the pursuit of short-run versus long-run aspirations (Hills & Hertwig, 2010). Finally, systematic differences in information search between single- and multi-play environments will help us to further understand how preferences, as revealed by choices, are further impacted by the

search that precedes them—a problem that generalizes to all tasks in which the actually experienced environment is a function of the organism's information search.

In the following, we first review pertinent empirical literature about expected utility in relation to single- and multi-play decisions. We then review findings on information search in decisions from experience, before describing how we link these lines of research.

### **1.1 Aspirations and the importance of expected utility in single- and multi-play decisions**

References to expected utility often invite one to say, subtly and under one's breath, 'long-run' expected utility. Some may argue that the addition of 'long-run' is redundant. Given the broad class of single-play decisions where expected utility does *not* immediately apply (Lopes, 1981), however, we would not agree. An offer to pay \$5 to play once a gamble that pays off \$100 with probability .1 and \$0 otherwise will leave the gambler poorer by \$5 nine times out of ten (Figure 1). This is true regardless of the amount of the non-zero payoff, be it \$100, \$1,000, or even \$100,000. However, the opportunity to play this gamble 100 times increases the probability of coming out ahead to above 50% (by 'coming out ahead,' we refer to the short-run aspiration of winning any non-zero amount). Anyone with a strict requirement of more than a non-zero return on their investment should avoid the single-play gamble, because in most realizations it will lead to losses. Recognizing this regularity, modern investment theory considers the time at which investment returns will be needed as a critical parameter in portfolio selection.



*Figure 1.* The influence of 1 play versus 100 plays on the probability distribution of outcomes at the end of play. Results for a single-play gamble costing 5 to play and promising an outcome of 100 with a probability of .1 and otherwise 0. The results for the multi-play gamble reflect the expected outcome per single play (each costing 5) of that gamble.

The importance of achieving a minimal aspiration and its role in explaining many choice anomalies has been well explored (Koop & Johnson, 2012; Lopes, 1996; see also Lopes & Oden, 1999). The key argument is that many of the mathematical prosthetics added to expected value theory (e.g., polynomial utility functions and subjective probability curves) are unnecessary if one considers that in many situations it may not be rational to pursue the expected value or long-run expected utility, but rather “the probability of coming out ahead”

(Lopes, 1981, p. 377)<sup>1</sup>. Indeed, studies investigating peoples' choices of single- and repeated-play gambles have found fewer violations of expected utility theory when people play repeatedly than when they play once (Camilleri & Newell, 2013; Keren, 1991; Keren & Wagenaar, 1987; Liu & Colman, 2009; Wedell & Böckenholt, 1990).

Behavioral ecologists have likewise discussed aspirations in relation to *risk sensitivity analysis* (i.e., the life and death consequences of decisions), noting that foragers adaptively change their preferences depending on the short-run requirements of their internal state. An animal that requires regular food to avoid starvation (e.g., a vampire bat's weight decays exponentially following a meal; Wilkinson, 1984) should favor those options that maximize the probability of acquiring food over those options that maximize long-run returns because the latter may leave it to starve in the meantime (e.g., Caraco, 1980; Houston & McNamara, 1999; Stephens, 1981, 2001; Stephens, Brown, & Ydenberg, 2007; see also Weber, Shafir, & Blais, 2004).

In the remainder of the manuscript, we conceptualize the contrast between single-play and multi-play in terms of aspirations, with short-run aspirations indicating an increased preference for the option that is most likely to come out ahead and long-run aspirations favoring the option offering the higher expected value. Yet, let us emphasize that aspirations are not the only way to conceptualize behavior in single-play and multi-play choice; there are numerous other accounts (Aloysius, 2007; Langer & Weber, 2001; Lopes, 1996; Tversky &

---

<sup>1</sup> The short-run aspiration of maximizing the chance of coming out ahead has sometimes been used interchangeably with the aspiration of maximizing some percentile of the outcome distribution (e.g., the median). Although both criteria would essentially produce identical predictions in our study, we focus on the aspiration of coming out ahead for two reasons. First, for two-outcome gambles as used in our study, the median is not well-defined. Second, the aspiration of maximizing the chance of coming out ahead corresponds more closely to the short-run criteria implemented in the literature on risk-sensitive foraging that we reference (e.g., Stephens, 2001).

Bar-Hillel, 1983; Wedell, 2011). Specifically, any mechanism explaining risk aversion, that is, the preference for the option with the lower variance, is under most circumstances also capable of explaining differences in choice (but not in search), even when the expected values of single- and multi-play scenarios are the same (as in Fig. 1). Such explanations include non-linear transformations of outcomes and non-linear transformations of probabilities (Kahneman & Tversky, 1979; see Wedell, 2011). The important point here is that the influence of both aspirations and non-linear weighting is diminished in multi-plays and the organisms can thus more ‘safely’ aim for the expected value.

In what follows, we turn to a choice environment in which—in contrast to Samuelson’s (1963) described option—the properties of the choice options (i.e., outcomes and probabilities) are not explicitly stated. Instead, people have to gather information prior to making a choice. This choice environment is representative of the myriad situations in which humans, and certainly other animals, need to make decisions based on experienced information samples rather than on symbolic descriptions of the world. Using this environment, we explore to what extent single- versus multi-play gambles affect search *and* choice when people make decisions from experience.

## **1.2 The relation between aspirations and information search**

Assuming that decision makers have different aspirations in single- and multi-play environments, the question is whether and how the process of information acquisition differs across these environments. If decisions and decision rules implementing these different aspirations require more or less information, then adaptive search strategies that meet such differential demands can foster better decisions by increasing efficiency. Consistent with the idea of adaptive information search, Hills and Hertwig (2010) found that specific information search policies in decisions from experience are associated with specific decision policies.

Individuals who took more samples and switched less often between options were more likely to choose options associated with maximizing expected utility, whereas individuals who took fewer samples and switched more frequently between options tended to choose options that won most of the time. Specifically, individuals who showed frugal search and avid switching appeared to accomplish this by comparing the promised return on the options between switches and choosing the option that won most of the time. Individuals who showed avid search and frugal switching, on the other hand, appeared to choose the option with the higher mean return computed from all collected samples. Though important for understanding the interplay of information search and choice, these findings are correlative. Consequently, they cannot discern between two possibilities: Do search policies sway later decision strategies, or do preselected decision strategies shape subsequent search policies?

Hills and Hertwig (2010) speculated that the correspondence between search and decisions could be driven by different aspirations (e.g., in line with risk sensitivity analysis in behavioral ecology). However, it is also plausible that the cognitive control of attention drives search, irrespective of top-down aspirations. Specifically, the sampling paradigm in research on decisions from experience has participants make a choice between two payoff distributions (hereafter referred to as options) after they have had the opportunity to explore (sample) them (Hertwig et al., 2004). A person may sample the outcomes \$0, \$0, \$0, and \$32 for one option and \$3, \$3, and \$3 for the other. Following a choice, the person would receive the value of one randomly drawn outcome for the payoff distribution he or she decided on. Using this sampling paradigm, Rakow, Demes, and Newell (2008, see also Ashby & Rakow, 2014) observed that total sample size and subsample sizes (samples between switches) were positively correlated with working memory span, a measure proposed to be associated with attentional control (e.g., Conway, Cowen, & Bunting, 2001; Kane & Engle, 2000).



Consequently, it is unclear to what extent cognitive control of attention, short-term versus long-term aspirations, or both drive search in this paradigm.

Apart from being a matter of theoretical interest, which mechanism—aspirations, cognitive control of attention, or both—drives search has important methodological implications. The key element in research on decisions from experience (involving the sampling paradigm) is that the experience one makes is a function of the amount of search. In particular, limited search carries the risk of systematically missing rare, but potentially consequential outcomes. If aspirations drive search, aspirations—that is, preference structures—determine not only what decision makers choose, but how they search for information prior to choice. This means, in turn, that their preference structure may not be uniquely identified on the basis of their choices, an issue with notable consequences that we will revisit in the discussion.

### **1.3 Testing the impact of short- and long-run aspirations on decisions from experience**

Most previous studies of decisions from experience (with the sampling paradigm; see Hertwig, in press) left it to the individual to pursue either long-run aspirations (thus aiming to maximize expected value in each problem by always choosing the option with higher expected value or, as a proxy, the option with the higher experienced mean) or short-run aspirations (thus aiming to maximize the probability of winning anything in each problem by always choosing the option that promises a greater chance of coming out ahead). In order to determine which aspirations people pursue spontaneously in the sampling paradigm, without explicit instructions, we compared it with conditions that highlight the short-run versus long-run consequences of decisions (similarly to Camilleri & Newell, 2013; Wedell & Böckenholt, 1990). Specifically, we informed participants that their final payoff depended either on a single, randomly chosen outcome from one of their chosen payoff distributions, multiplied by

100 (*single-play condition*), or on 100 random draws from one of their chosen distributions (*multi-play condition*). Except for these payoff instructions, both conditions were identical. Because search in the sampling paradigm is entirely self-directed, with participants free to sample from the payoff distributions for as long as and in whatever order they like, this set-up means that—apart from the influence of sampling—both conditions rest on the same choice environment. As in many real-life decisions, it is left to participants to infer the consequences of a single play or multiple plays. This set-up allows us to directly evaluate and compare patterns of information search and choices across a total of three implementations of the sampling paradigm: the single-play condition, the multi-play condition, and the standard implementation (where a single draw from *each* chosen payoff distribution determines the final payoff).

Furthermore, we designed decision problems with a structure often employed in the decisions-from-experience literature, requiring a choice between a risky option (with two outcomes) and a safe option. As we demonstrate shortly, these problems have the property that the number of samples needed to detect the option that promises the larger probability of coming out ahead is *less* than the number of samples needed to detect, with the same precision, the option with the larger mean.

We expect that assigning short- and long-run consequences to otherwise identical choice environments will lead to individuals adapting their information search and choice policies to their aspirations. In other words, people may not only be adaptive decision makers (Payne, Bettman & Johnson, 1993), but also adaptive information searchers. Furthermore, we predict that aspiration-induced differences in search will prompt systematically different experiences of identical options. Specifically, in the single-play condition, relative to the multi-play condition, fewer people will experience the rare event—because of smaller sample

sizes. By extension, we predict that participants in the multi-play condition, relative to the single-play one, will more likely choose the option with the higher expected value.

Furthermore, we expect that behavior in the standard sampling paradigm will lie somewhere in between that of respondents in the single-play and multi-play conditions, both in search and choice, thus reflecting the interindividual heterogeneity of potential short- and long-run strategies previously observed (Hills & Hertwig, 2010).

Finally, in order to evaluate the potential role of attentional control on information search, we also measured each participant's operation span. Operation span is a complex working memory span measure that taps into a person's ability to store and retrieve a sequence of individual tokens over intermittent distractor tasks. Alternative to or in addition to short- and long-run aspirations, attentional control may determine search. Based on previous findings (Rakow et al., 2008), we expect higher operation spans to be associated with taking more samples and with fewer switches between the options. Moreover, one may speculate that attentional control is also linked to choice. The short- and long-run choice policies suggested by Hills and Hertwig (2010) are likely to require different levels of cognitive effort. Specifically, the short-run policy may be less cognitively demanding, requiring the simple tallying, across multiple comparisons, of the number of wins for each option. The long-run policy of choosing the option with the higher mean, on the other hand, requires the integration of all observed information in combination with the weighting of the returns by the total sample size; both of these processes may require more attentional control. Consequently, the coupling of search and choice may also be caused and explained by individual differences in attentional control capacities (vs. differences in short-run and long-run aspirations).

## **2. Method**

## 2.1 Participants

We collected data from 124 students of the University of Basel, with a mean age of 24 years. Participants were rewarded with either course credit or a fixed payment of approximately \$13. In addition, participants received a performance-based bonus as a result of their choices.

## 2.2 Materials

We designed 12 decision problems (Table 1). The two options within each problem offered the opportunity to maximize either the long-run expectation (higher mean or expected value) or the probability of coming out ahead (higher median). Each problem presented a choice between a high-outcome rare event ( $p \leq .15$ ) in the higher expected value option (otherwise 0) and a small but relatively secure outcome ( $p \geq .7$ ) in the lower expected value option (otherwise 0). All problems share the property that more samples are required to spot the option with the higher mean than the option with the higher probability of coming out ahead. To demonstrate this, we simulated 10,000 decisions for each problem and determined how many samples would be needed to identify the option with the higher mean versus the higher chance of coming out ahead, given some level of precision. Identifying the latter with a probability of, for instance, at least 80% requires a much smaller sample than identifying the higher mean option (on average, about 4 vs. 34 draws per option). Of course, we do not expect our participants to know in advance the underlying distributions; however, they may rapidly develop an understanding of these distributions once sampling begins. Sample sizes will reflect this grasp.

Table 1. Decision problems employed in the three conditions

Problem	<i>H</i>	<i>L</i>
1	92 <sup>1</sup> with $p = .05$	3 with certainty
2	34 with $p = .05$	1 with certainty
3	120 with $p = .05$	5 with $p = .70$
4	44 with $p = .05$	2 with $p = .70$
5	70 with $p = .10$	4 with certainty
6	16 with $p = .10$	1 with certainty
7	54 with $p = .10$	4 with $p = .75$
8	23 with $p = .10$	2 with $p = .75$
9	35 with $p = .15$	3 with certainty
10	21 with $p = .15$	2 with certainty
11	48 with $p = .15$	5 with $p = .80$
12	9 with $p = .15$	1 with $p = .80$

Note: *H* = option with the higher expected value (as calculated by probability  $\times$  monetary value); *L* = option with the lower expected value.

<sup>1</sup> In order to provide identical incentives across conditions, we matched the expected returns across conditions by multiplying each randomly drawn outcome in the standard condition by a factor of 6.

In order to prevent participants from inferring that the option with the rare and consequential event always promised the higher expected value, we intermixed four problems with a different structure (Appendix, Table A1), resulting in a total of 16 decision problems. All were included in our analysis, because the predictions for the search and decision strategies are qualitatively independent of the structure of the decision problems. To measure individuals' working memory capacity, we used the automated version of the *operation span task* (Unsworth, Heitz, Schrock, & Engle, 2005). This computer-based task measures people's ability to remember sets of letters (e.g., E A D) that appear, one letter at a time, following simple math problems (e.g.,  $[1 \times 2] + 1 = ?$ ). The operation span score was determined as the sum of all correctly recalled letter sets (over a series of sets with lengths ranging from 3 to 7).

### 2.3 Procedure

Problems were presented on a computer screen. Participants were randomly assigned to one of the three conditions. In the multi-play condition, they were instructed that their payoff would be determined by randomly selecting one of their final choices and then taking 100 random draws from the selected option (e.g., 100 draws from selected option  $H$  in Problem 1: 92 with probability .05; Table 1). In the single-play condition, participants were instructed that one of their chosen options would be randomly selected; a single random draw from this option would then be taken, and the resulting outcome would be multiplied by a factor of 100. This procedure renders the magnitudes of the expected values in the single- and multi-play conditions identical. Finally, the *standard* condition implemented the payoff modality used in past studies (e.g., Hertwig et al., 2004). Specifically, a random draw from each chosen option across the 16 problems determined the payoff. Because this meant that 16 draws were incentivized in the standard condition, relative to 100 draws in the other two conditions, we matched the expected returns across conditions by multiplying each randomly drawn outcome in the standard condition by a factor of 6.

Before participants turned to the 16 decision problems (presented in a random order), they worked on three practice trials. For each problem, they were able to sample from the two options as extensively and in whatever fashion they liked. Once search was terminated, they proceeded to their final choice by clicking a button. The operation span measure was administered once all choices were completed. At the end of the experiment, participants received a bonus as a result of their choices, paid out one-to-one in accordance with the conditions' payoff scheme.

For the purpose of data analysis, we set the threshold that participants had to sample at least once from both options in at least half of the problems; five participants (of 124; 4%)

failed to meet this criterion and were removed from the analysis. Additionally, all trials in which a person sampled only a single option were removed. The following analyses are thus based on 95% of all trials provided by 119 participants.

### 3. Results

#### 3.1 The influence of single-play and repeated-play on search and switching

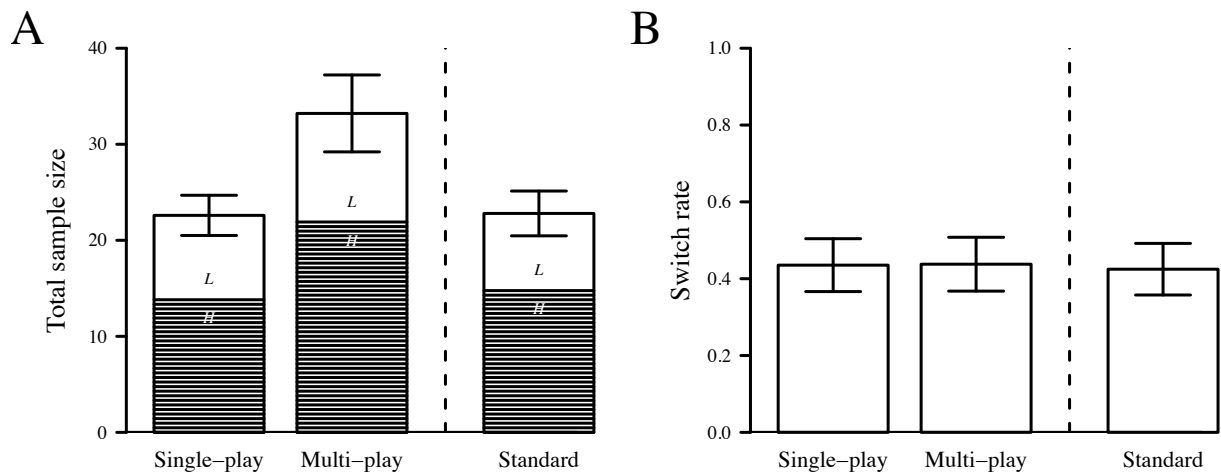
As Figure 2 shows, amount of search in the repeated-play condition was very different from that in the single-play condition. On average, participants in the former condition took about 10.6 samples more than those in the latter condition (total sample size:  $t[77] = 2.37, p = 0.02, d = 0.53$ ).<sup>2</sup> This difference was mainly driven by larger samples from the risky or riskier option (option *H* in Table 1) accounting for 7.1 of the additional 10.6 samples,  $t(77) = 2.13, p = .04$ . However, there also was more extensive search in the safe or safer option (option *L* in Table 1;  $t[77] = 2.43, p = .02$ ). Because participants in all conditions took, on average, about two samples from option *H* for every one sample from option *L* ( $n_H/n_L = 1.75\text{--}2.05$ ), it appears that participants in the multi-play condition increased their search effort about equally for both options. What about search in the standard sampling condition? Sample size was different from that achieved in the multi-play condition,  $t(77.1) = 2.26, p = .03$ , but almost identical to that in the single-play condition,  $t(78) = .1, p = .95$ .

Previous studies found that sample size and switch rate were correlated in the standard sampling paradigm (Hills & Hertwig, 2010; Rakow et al., 2008). Moreover, Hills and

---

<sup>2</sup> Global tests of significance were omitted due to clear hypotheses for the pairwise group comparisons. All reported  $t$  values were derived from mixed effects analyses predicting the outcome variable on the problem level while controlling for the subject variable via the inclusion of a random intercept. Tests were performed using the statistical software R (R Development Core Team, 2008) and the packages lme4 and lmerTest. Specifically, Gaussian linear models were estimated using REML and Satterthwaite's approximation for degrees of freedom, the default method in lmerTest. The effect size  $d$  is a standardized measure, and  $d = .2, .5$ , and  $.8$  denote small, medium, and large effects, respectively (Cohen, 1988).

Hertwig (2010) found that switch rate was inversely correlated with the choice of the option with the higher expected value. These observations invite the question as to whether decision strategies govern switch rate. Our results suggest they do not. The number of switches per sample was not significantly different between conditions,  $t(77) = 0.33$ ,  $p = .75$  (see Fig. 2B). This finding contradicts Hills and Hertwig's (2010) suggestion that frequent switching may be caused by the aspiration of short-run maximization. Moreover, switch rate in the standard paradigm was not statistically different from that in the multi-play (switch rate:  $t[77] = 0.06$ ,  $p = .95$ ) or single-play condition (switch rate:  $t[78] = 0.23$ ,  $p = .77$ ). To summarize, the single-versus multi-play instructions markedly affected sample size; the switch rate, in contrast, appeared to be less sensitive to the difference in aspiration.



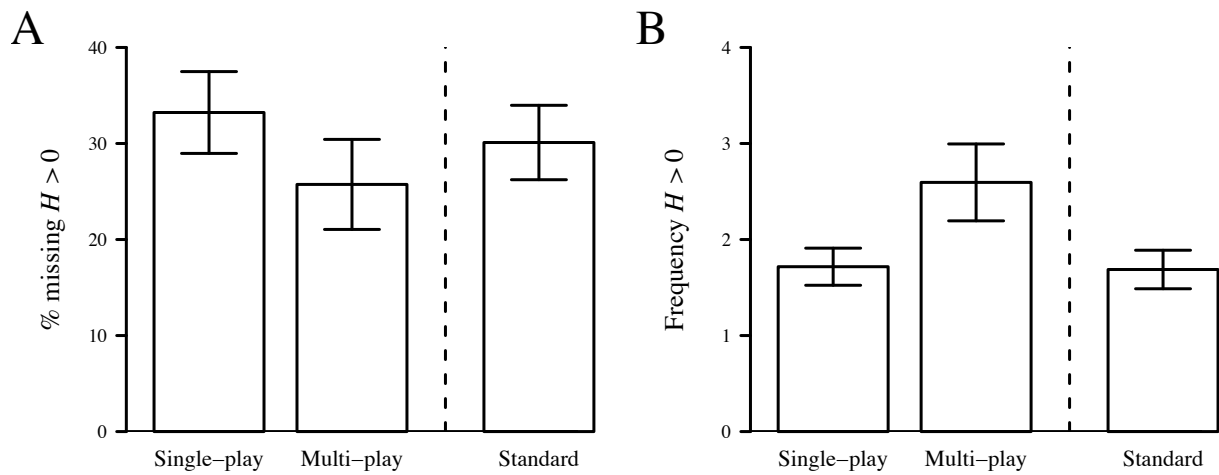
*Figure 2.* Total sample size (A) and switch rate (B), separately for the three conditions: single-play, multi-play, and the standard sampling paradigm. In panel A, bars are further split into samples taken from the *H* and *L* options (see Table 1). Error bars represent the standard error of the mean.

### 3.2 How single- versus multi-play shape individuals' experience



One key characteristic of the sampling paradigm is that an individual's sample size inevitably shapes his or her experience of the events' probabilities. In particular, rare events are often not encountered when sample sizes are small; and even if they are observed, the number of people who experience them less frequently than expected exceeds the number who experience them more frequently than expected (as a consequence of the skewness of the binomial probability distribution for small  $ns$  and small  $ps$ ; see Hertwig et al., 2004). The large difference in sample size observed for the multi-play and single-play conditions is thus likely to translate into different experiences: Specifically, participants in the multi-play condition are, *ceteris paribus*, more likely to be *cognizant* of the rare positive outcome in option  $H$  than are participants in the single-play condition; furthermore, the former can be expected to have *experienced* the rare event more often than the latter. Is this indeed the case?

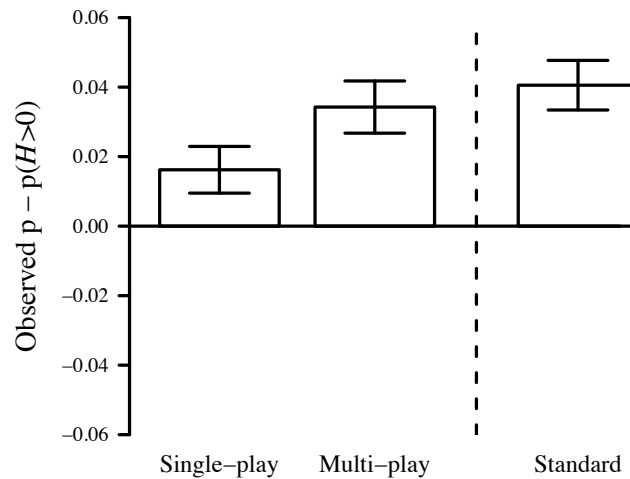
In terms of awareness of rare events, our findings showed that participants in the single-play condition were 1.3 times more likely to miss the rare event than were participants in the multi-play condition (33% vs. 26%). Although this difference was in the expected direction (in light of the different sample sizes; see Figure 2A), it was not significant (logit link:  $z = 1.42, p = .16$ ). But did the frequency with which people experienced a rare event (given that it was encountered once) differ between conditions? Figure 3A shows that the rare event was encountered 1.5 times more often in the multi-play condition than in the single-play condition,  $t(77) = 2, p = 0.049, d = .45$ .



*Figure 3.* Percentage of participants who missed (A) and frequency of encountering (B) the rare positive outcome in option  $H$ , separately for the three conditions: single-play, multi-play, and standard sampling paradigm. Analysis based on the 12 decision problems involving a rare positive event in option  $H$ . Error bars represent the standard error of the mean.

Why were people in the single-play condition not markedly more likely to miss the rare event than people in the multi-play condition? One possible explanation relates to optional stopping. To the extent that sample size (i.e., number of draws) is determined at the outset of the sampling process, the binomial probability distribution governs the sampling process; furthermore, it implies a smaller chance to observe the rare event with smaller sample sizes (see Hertwig et al., 2004, and Hertwig & Pleskac, 2010). Alternatively, however, stopping may be controlled by the actually experienced outcomes, thus rendering the binomial distribution an inappropriate model (see, e.g., Berger & Berry, 1988, for a discussion of optional stopping in statistical inference). Our data indicate that experience matters for the decision to stop. Figure 4A plots the differences between the observed probabilities of the rare events and their true probabilities in the multi-play,  $t(37.2) = 4.66$ ,  $p < 0.001$ , and single-play condition,  $t(39.2) = 2.4$ ,  $p = 0.02$ . In both conditions, people experienced the rare event more

frequently than expected, consistent with outcome-dependent stopping—in other words, people appear to have stopped shortly after observing a rare event.



*Figure 4.* Observed probability minus true probability of the rare positive outcome in option  $H$ , separately for the three conditions: single-play, multi-play, and standard sampling paradigm. Analysis based on the 12 decision problems involving a rare positive event in option  $H$ . Relative frequencies are displayed in comparison to the expectation, i.e., the respective true probability. Error bars represent the standard error of the mean.

Fewer samples in total and fewer observations of the rare event in option  $H$  relative to the multi-play condition (Fig. 3A; multi-play vs. standard:  $t[77.2] = 2.04$ ,  $p = 0.04$ ) suggest that participants in the single-play and the standard sampling condition mustered experiences that were similar. This did not hold for every dimension, however. Participants in the standard condition experienced rare events more often than those in the single-play condition (see Fig. 3B;  $t[77.2] = 2.04$ ,  $p = 0.04$ ). They also experienced a larger mean difference in returns in favor of option  $H$  (see Fig. 4B;  $t[75.7] = 2.93$ ,  $p < .001$ ,  $d = 0.65$ ). Thus, although the single-

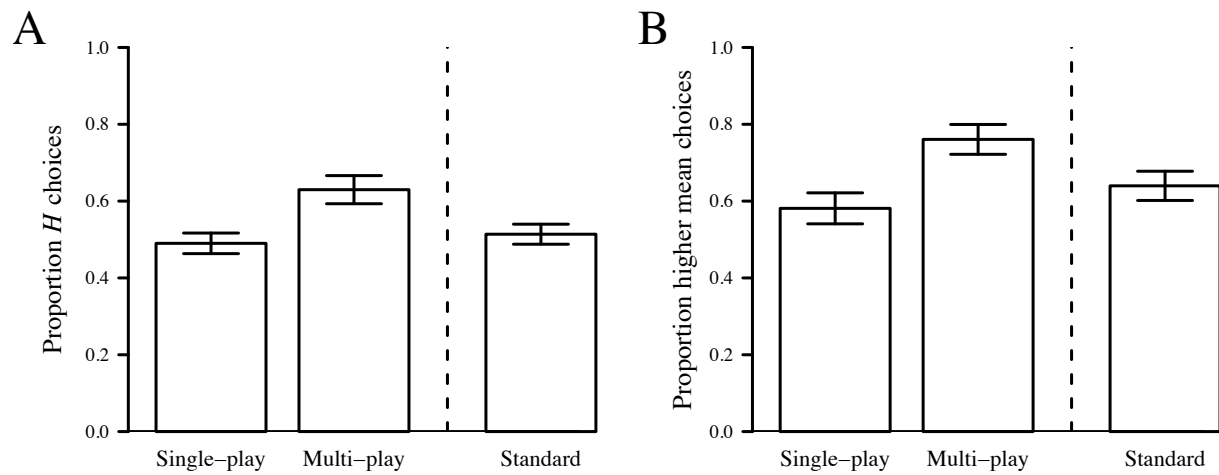
play and the standard condition were alike in terms of sample size and switching, the samples on which they based their choices were not identical.

### 3.3 The influence of single-play and repeated-play on choice

We started out, among other hypotheses, by predicting that induction of long-run aspirations would lead to favoring the option with the higher expected value. Do our data support this hypothesis? Figure 4 plots the proportion of choices of option  $H$ . Consistent with this hypothesis, the proportion of choices of option  $H$  in the multi-play condition was 63%, substantially higher than the 49% observed in the single-play condition ( $z = 3.16, p = 0.002, OR = 1.92$ ). Relatedly, the proportion of choices of option  $H$  in the standard condition was 51%, which was not statistically different from that observed in the single-play condition ( $z = 0.67, p = 0.5$ ), but was different from that observed in the multi-play condition ( $z = 2.66, p = 0.01$ ).<sup>3</sup> These similarities and differences thus suggest that long-run aspirations are conducive to expected-value maximization and that respondents appear to perceive the standard paradigm as a one-shot decision, notwithstanding the opportunity to aggregate choices across the sequence of decision problems in the experiment (Read, Loewenstein, & Rabin, 1999).

---

<sup>3</sup> A mixed effects regression indicated that  $H$  choices were not influenced by the presence or absence of certainty in the  $L$  option ( $z = 1.53, p = 0.13$ ). Because none of the following analyses were influenced by a comparison of safe and risky options, the following results are collapsed across both.



*Figure 5.* Proportion of choices of the higher expected value option  $H$  (A) and choices consistent with the higher experienced sample mean (B), separately for the three conditions: single-play, multi-play, and standard sampling paradigm.

However, there is an important caveat to this interpretation. As spelled out before, differences in choices between single- and multi-play conditions could also be due to the different information people experienced. One and the same payoff distribution, once filtered through experience, can take on many different ‘phenotypes.’ It is possible that respondents in both conditions maximize the same quantity, but that the quantity is the *experienced* mean reward, rather than the expected value. In other words, all the difference in choice might reside in the difference in the sampled information (see Fig. 2, 3, and 4) and thus in the experienced phenotype rather than in different proclivities to maximize. In order to test this possibility, we next calculated the proportion of choices of the option with the higher experienced mean (in those cases where participants experienced the rare event; 71% of cases) while accounting for the observed difference in means. Figure 4B shows the results. Individuals in the multi-play condition continued to be much more likely to choose options with the higher experienced mean than were individuals in the single-play condition ( $z = 3.3$ ,

$p < .001$ ,  $OR = 2.67$ ). This effect was not reduced by the inclusion of the actual difference in means between the options ( $z = 2.8$ ,  $p = 0.01$ ,  $OR = 2.68$ ). This suggests that differences in choices are not a mere function of sampling error (see Hertwig & Erev, 2009). Moreover, in further support of the notion that participants tend to view the standard paradigm as a short-run scenario, the standard condition did not differ from the single-play condition ( $z = 0.76$ ,  $p = 0.44$ ), but did differ from the multi-play condition ( $z = 2.54$ ,  $p = 0.01$ ).

Finally, let us unpack one finding reported in Figure 4. Although choices in the single-play condition were much less likely to maximize expected value (4A) and experienced mean (4B) than were those in the multi-play condition, people still chose the higher mean option in 58% (versus 76%) of cases. This is of course not perfectly compatible with the notion that people in the single-play option tended to maximize the chances of coming out ahead (i.e., the median reward). However, people in the single-play condition also chose the option with the higher experienced chance of coming out ahead in 56% of cases (in 23% of cases, the options with the higher experienced mean and the higher chance of coming out ahead were identical). In contrast, people in the multi-play condition did so only in 44% of cases. One interpretation of this finding is that people in the single-play condition find themselves halfway between the two goals, with some betting on the rare but attractive gain, and others trying to come out ahead. In the multi-play environment, in contrast, the predominant course of action is maximization of the experienced mean.

### **3.4 The role of working memory capacity**

Based on a previously observed association between search and working memory (Rakow et al., 2008), we hypothesized that attentional control may serve as plausible explanation for the dependency between search and choice. To test this relationship, we measured participants' operation span and evaluated its association with sample size, switch rate, choices of the

higher expected value option  $H$ , and choices of the option with the higher experienced mean. Table 2 shows the results of independent mixed effects analysis predicting these variables by operation span score. None of the effects were significant, suggesting that—if at all—attentional control plays a limited role in explaining search and choice in the sampling paradigm.

*Table 2: Mixed-effects regression of search and choice on operation span*

	Sample size	Switch rate	$H$ choices	Higher mean choices
Single-play	$b = 3.14,$ $p = 0.14$	$b = 0.03,$ $p = 0.2$	$b = 0.01,$ $p = 0.6$	$b = 0.02,$ $p = 0.66$
Multi-play	$b = 2.65,$ $p = 0.52$	$b = 0.01,$ $p = 0.82$	$b = 0.05,$ $p = 0.19$	$b = 0.04,$ $p = 0.29$
Standard	$b = 0.25,$ $p = 0.91$	$b = 0.03,$ $p = 0.32$	$b = 0.02,$ $p = 0.41$	$b = 0.02,$ $p = 0.6$

*Note:* Estimates ( $b$ ) correspond to the change in the respective variable given a change of one standard deviation in the operation span score. Higher mean choices include only those choices where the rare event was observed.

#### 4. General discussion

Following up on Samuelson's anecdotal observation (1963), we investigated the suggestion that people making decisions from experience may choose differently when playing a gamble once versus multiple times (Lopes, 1996; Wedell, 2011). Using the sampling paradigm, we found differences in both choice and information search between single- and multi-play conditions. In the multi-play condition, individuals sampled more and were more likely to choose options with the higher expected values than did participants in the single-play condition. These differences were not mediated by experiencing different choice environments (and sampling error; see Fox & Hadar, 2006; Hertwig et al., 2004). Instead, they appear to stem from changes in decision strategy that were foreshadowed by changes in

information search, an outcome consistent with the idea of adaptive information search as proposed by Hills and Hertwig (2010).

#### **4.1 Implications for single- and multi-play choices**

Our results add to the descriptive debate on single- and multi-play choices. The normative debate—i.e., whether people should have stable preferences across single- and multi-play situations—has cooled off, but the discussion of how to best conceptualize the psychological processes involved in single- and multi-play situations is still ongoing. The first of two major positions is exemplified by Lopes' security potential and aspiration theory (SP/A; Lopes & Oden, 1999; see also Wang & Johnson, 2012; Payne, 2005) and proposes that (at least) two separate processes are executed in sequence: First, prospects are qualitatively compared against some aspiration level. When, and only when, the aspiration level is satisfied, the individual engages in a second, more systematic valuation of the prospect. If this is the case, the evaluation process of single- and multi-play situations could differ markedly, because multi-play prospects are more likely to surpass the aspiration level and trigger a systematic valuation than are single-play prospects (Wedell, 2011). The second position, exemplified by cumulative prospect theory (CPT; Tversky & Kahneman, 1992), proposes that just one process operates for single- and multi-play prospects alike: Outcomes and probabilities, once acquired and before being integrated into a single utility value, undergo non-linear transformations that allow single-play and multi-play versions of the same prospect to yield different utilities. Without making additional assumptions, this position implies identical evaluation processes for single-play and multi-play scenarios. Previous investigations using fully described single- and multi-play prospects have found evidence for differences in the behavioral patterns of choice and information acquisition (Joag, Mowen, & Gentry, 1990; Su et al., 2013), as well as in post-hoc verbal reports (Wedell & Böckenholt, 1990). Our



investigation using decisions from experience adds to this debate by showing that the amount of information sampled prior to choice and the resulting experience varies in response to single- and multi-play instructions. Consistent with theoretical accounts assuming multiple processes, this finding suggests that the valuation process indeed differs between single- and multi-play choices.

#### **4.2 Implications for search and choice in decisions from experience**

Our findings also provide new insights into the psychology underlying the standard sampling paradigm often used in recent research on decisions from experience. Behavior in this paradigm most resembled that observed in the single-play environment. This finding may be somewhat surprising, given that the standard condition, like the multi-play condition, offered multiple draws—one draw for each of the 16 choices. Participants in the standard condition could thus also aggregate the risk by bracketing the choices together (see Read et al., 1999). The results, however, suggest that participants evaluated each choice individually. This is consistent with previous research using decisions from description showing that people usually tend to segregate prospects when the cumulative nature of multiple prospects is not made apparent (Redelmeier & Tversky, 1992). Thus, one explanation may be that participants did not realize that they could aggregate the risk across their multiple choices (see DeKay & Kim, 2005, for the role of perceived fungibility in multi-play choices). An alternative explanation may reside in computational complexity of the respective choice strategies. Short-run maximization is likely to be simpler—in terms of its computational and memory demands—than long-run maximization. For options like those presented here, an individual could simply count the number of zeroes that occur during sampling and choose the option with fewer zeroes (assuming approximately equal sample sizes per option). Computing the

means, in contrast, requires additional steps that involve, at the least, adding non-zero numbers (Hau, Pleskac, Kiefer, & Hertwig, 2008).

The finding that choice and search behavior in the standard condition resembled behavior in the single-play condition is of particular relevance to the discussion of as-if patterns of underweighting of rare events, and to what has been termed the description–experience gap (Hertwig & Erev, 2009). In contrast to the regularity of low-probability events tending to be overweighted in decisions from description (Hertwig & Erev, 2009; Tversky & Kahneman, 1992), it has been inferred from choices that rare events tend to receive *less* weight than they deserve in decisions from experience (Camilleri & Newell, 2011; Hertwig et al., 2004). Studies using yoked experiences and direct probability judgments by participants have revealed that such underweighting also occurs when the subjective representation of the prospect accurately reflects its objective properties (Camilleri & Newell, 2009; Ungemach, Chater, & Stewart, 2009). The use of short-run strategies that maximize the chance of coming out ahead by inherently ignoring rare events may offer a new angle from which to address the persistent puzzle of why such underweighting occurs.

Our results also inform previous hypotheses regarding frugal search efforts in decisions from experience (see Hertwig & Pleskac, 2010). Two main arguments have been invoked to explain the relatively modest sample sizes observed. First, working memory limitations may provide a natural stopping rule for search (Hertwig et al., 2004; Rakow et al., 2008; see also Kareev, 2000). Second, Hertwig and Pleskac (2010) have suggested that small samples amplify the difference between the expected earnings associated with the different payoff distributions, thus making the options in question more distinct and, consequently, choice easier. The present investigation identifies a third contributing factor: Search is a function of

people's aspirations, and if many or even most people tend to pursue short-run aspirations (in the sampling paradigm), frugal search follows naturally.

Further, our results suggest that different components of explorative behavior may be under different control processes. The lack of a difference in switch rate across all conditions appears to indicate that this property is not part of a participant's top-down aspiration level and associated decision strategy. Rather, switching may be under the control of more implicit processes, such as those associated with working memory (Hills & Pachur, 2012; Rakow et al., 2008; see also Hills, Todd, & Goldstone, 2010). Yet the complete independence between our measure of working memory and both search and choice variables throws into question working memory's potential role as a stopping rule as well as its previously suggested role (Rakow et al., 2008) as a common cause for search and choice policy (for similar findings, see Wulff, Hills, & Hertwig, 2014). One reason for the absence of a link may lie in the high complexity and difficulty of the operation span task relative to the simple digit span task used by Rakow et al. (2008). Yet, it could also mean that models for decision from experience that simplify information integration and choice, thereby taxing working memory less, may be good candidate models for experienced-based decision making. One particular class, often labeled as associative learning models (see Hertwig, *in press*; Sutton & Barto, 1998), assumes a continuous updating of a single utility value per option and thus requires the storage of far less information. Similarly, the maximization of the experienced mean could be achieved by recruiting a simple heuristic such as the natural mean heuristic (Hertwig & Pleskac, 2010). Future investigations should explore how short- and long-run aspirations may be captured in choice models based on an associative learning mechanism. In this context, attention should also be paid to the diversity of executive functions (Miyake et al., 2000), of which only some may be involved in experienced-based risky choice.

Our results also shed light on an aspect of information search in decisions from experience that has largely been overlooked, namely, optional stopping. Deliberations into the statistical effects of small samples have often assumed that search is randomly terminated or is terminated once a preplanned size is reached (Fox & Hadar, 2006; Gonzales & Dutt, 2011; Hertwig et al., 2004; Hertwig & Pleskac, 2010). Our investigation suggests that termination of search may also be subject to strategic concerns (for related findings, see Lejarraga, Hertwig, & Gonzales, 2012; Phillips, Hertwig, Kareev, & Avrahami, 2014).

#### **4.3 The problem of inferring risk preference in decisions from experience**

Last but not least, our findings highlight a thorny inference problem concerning risk preferences in experienced-based choices. Following the revealed preference approach (Samuelson, 1938), researchers often infer an individual's preference directly from her choices in described and stable choice environments (decisions from description). Decision problems employing stated probabilities can easily be tailored to make different risk preferences discernible (e.g., Holt & Laury, 2002). In decisions from experience, though, inferring risk preference from choice is much more problematic. Due to variable sample sizes and the random composition of samples, it is often the case that no two individuals face identical decision problems. For this reason, it has been convincingly argued that the individual-specific experienced choice environment—and not the objective choice parameters (outcomes and probabilities)—is the appropriate foundation for inferring the individual's risk preference (Fox & Hadar, 2006). Our findings, however, suggest that even inferring preferences contingent on experience is problematic. For illustration, consider a person with a *strong* preference for coming out ahead in the short term. She takes relatively few samples in each problem. Consequently, she may be faced with a decision between an apparently safe and modest positive outcome and an apparently safe zero outcome. This 'trivial' (dominated)

choice reveals little about this person's risk preference. Now, consider a person with a *weak* preference for coming out ahead in the short term. She may sample a bit more than the first person, and even encounter the rare but attractive outcome. Because of her preference, she decides against the option offering this dicey but attractive outcome. This person will be 'revealed' to be risk averse, whereas the other appears, if anything, to be risk neutral. Of course, this is a constructed example (ignoring, among other factors, the role of optional stopping), but it illustrates a simple but consequential point: In environments that people experience and 'construct' through active sampling, inferences from choice to preference are problematic because the experienced environment can arise from the preference or aspiration level itself. Depending on which environment emerges, choices may or may not be informative about the underlying preferences or aspirations. The good news, however, is that decisions from experience paradigms offer an *observable* psychological dimension that appears to afford researchers another window onto preferences or aspirations: the appetite for information (see also Denrell, 2007).

### References

- Aloysius, J. A. (2007). Decision making in the short and long run: Repeated gambles and rationality. *British Journal of Mathematical and Statistical Psychology*, 60, 61–69.
- Ashby, N. J., & Rakow, T. (2014). Forgetting the past: Individual differences in recency in subjective valuations from experience. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 40(4), 1153–1162.
- Berger, J. O., & Berry, D. A. (1988). Statistical analysis and the illusion of objectivity. *American Scientist*, 76, 159–165.
- Camilleri, A. R., & Newell, B. R. (2009). The role of representation in experience-based choice. *Judgment and Decision Making*, 4(7), 518–529.
- Camilleri, A. R., & Newell, B. R. (2011). When and why rare events are underweighted: A direct comparison of the sampling, partial feedback, full feedback and description choice paradigms. *Psychonomic Bulletin & Review*, 18(2), 377–384.
- Camilleri, A. R., & Newell, B. R. (2013). The long and short of it: Closing the description–experience “gap” by taking the long-run view. *Cognition*, 126, 54–71.
- Caraco, T. (1980). On foraging time allocation in a stochastic environment. *Ecology*, 61, 119–128.
- Cohen, J. (1988): Statistical power analysis for the behavioral sciences (2nd ed.). Hillsdale, NJ: Erlbaum.
- Conway, A. R. A., Cowan, N., & Bunting, M. F. (2001). The cocktail party phenomenon revisited: The importance of working memory capacity. *Psychonomic Bulletin & Review*, 8, 331–335.
- DeKay, M. L., & Kim, T. G. (2005). When things don’t add up: The role of perceived fungibility in repeated-play decisions. *Psychological Science*, 16(9), 667–672.

- Denrell, J. (2007). Adaptive learning and risk taking. *Psychological Review*, 114(1), 177–187.
- Fox, C. R., & Hadar, L. (2006). “Decisions from experience” = sampling error + prospect theory: Reconsidering Hertwig, Barron, Weber & Erev (2004). *Judgment and Decision Making*, 1, 159–161.
- Gonzalez, C., & Dutt, V. (2011). Instance-based learning: Integrating sampling and repeated decisions from experience. *Psychological Review*, 118(4), 523–551.
- Hau, R., Pleskac, T.J., Kiefer, J., & Hertwig, R. (2008). The description–experience gap in risky choice: The role of sample size and experienced probabilities. *Journal of Behavioral Decision Making*, 21, 493–518.
- Hertwig, R. (in press). Decisions from experience. In G. Keren & G. Wu (Eds.), *Blackwell handbook of decision making*. Oxford, UK: Blackwell.
- Hertwig, R., Barron, G., Weber, E. U., & Erev, I. (2004). Decisions from experience and the effect of rare events in risky choice. *Psychological Science*, 15, 534–539.
- Hertwig, R., & Erev, I. (2009). The description–experience gap in risky choice. *Trends in Cognitive Sciences*, 13, 517–523.
- Hertwig, R., & Pleskac, T. J. (2010). Decisions from experience: Why small samples? *Cognition*, 115, 225–237.
- Hills, T. T., & Hertwig, R. (2010). Information search in decisions from experience: Do our patterns of sampling foreshadow our decisions? *Psychological Science*, 21, 1787–1792.
- Hills, T. T., & Pachur, T. (2012). Dynamic search and working memory in social recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 38(1), 218.
- Hills, T. T., Todd, P. M., & Goldstone, R. L. (2010). The central executive as a search process: Exploration and exploitation in generalized cognitive search processes. *Journal of Experimental Psychology: General*, 139, 590–609.

- Holt, C. A., & Laury, S. K. (2002). Risk aversion and incentive effects. *The American Economic Review*, 92(5), 1644–1655.
- Houston, A. I., & McNamara, J. M. (1999). *Models of adaptive behavior: An approach based on state*. Cambridge, UK: Cambridge University Press.
- Joag, S. G., Mowen, J. C., & Gentry, J. W. (1990). Risk perception in a simulated industrial purchasing task: The effects of single versus multi-play decisions. *Journal of Behavioral Decision Making*, 3, 2, 91–108.
- Kahneman, D., & Lovallo, D. (1993). Timid choices and bold forecasts: A cognitive perspective on risk taking. *Management Science*, 39, 1, 17–31.
- Kane, M. J., & Engle, R. W. (2000). Working memory capacity, proactive interference, and divided attention: Limits on long-term memory retrieval. *Journal of Experimental Psychology: Learning, Memory, Cognition*, 26, 333–358.
- Kareev, Y. (2000). Seven (indeed, plus or minus two) and the detection of correlations. *Psychological Review*, 107(2), 397–402.
- Keren G. (1991). Additional tests of utility theory under unique and repeated conditions. *Journal of Behavioral Decision Making*, 4, 297–304.
- Keren, G., & Wagenaar, W. A. (1987). Violation of utility theory in unique and repeated gambles. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 13, 387–391.
- Koop, G. J., & Johnson, J. G. (2012). The use of multiple reference points in risky decision making. *Journal of Behavioral Decision Making*, 25, 49–62.
- Langer, T., & Weber, M. (2001). Prospect theory, mental accounting, and differences in aggregated and segregated evaluation of lottery portfolios. *Management Science*, 47(5), 716–733.



- Lejarraga, T., Hertwig, R., & Gonzalez, C. (2012). How choice ecology influences search in decisions from experience. *Cognition*, 124(3), 334–342.
- Liu, H. H., & Colman, A. M. (2009). Ambiguity aversion in the long run: Repeated decisions under risk and uncertainty. *Journal of Economic Psychology*, 20, 277–284.
- Lopes, L. L. (1981). Decision making in the short run. *Journal of Experimental Psychology: Human Learning and Memory*, 7, 377–385.
- Lopes, L. L. (1996). When time is of the essence: Averaging, aspiration, and the short run. *Organizational Behavior and Human Decision Processes*, 65, 179–189.
- Lopes, L. L., & Oden, G. C. (1999). The role of aspiration level in risky choice: A comparison of cumulative prospect theory and SP/A theory. *Journal of Mathematical Psychology*, 43, 286–313.
- Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., & Wager, T. D. (2000). The unity and diversity of executive functions and their contributions to complex “frontal lobe” tasks: A latent variable analysis. *Cognitive Psychology*, 41(1), 49–100.
- Montgomery, H., & Adelbratt, T. (1982). Gambling decisions and information about expected value. *Organizational Behavior and Human Performance*, 29, 39–57.
- Payne, J. W. (2005). It is whether you win or lose: The importance of the overall probabilities of winning or losing in risky choice. *Journal of Risk and Uncertainty*, 30(1), 5–19.
- Payne, J. W., Bettman, J. R., & Johnson, E. J. (1993). *The adaptive decision maker*. Cambridge, UK: Cambridge University Press.
- Peköz, E. A. (2002). Samuelson’s fallacy of large numbers and optional stopping. *Journal of Risk and Uncertainty*, 69, 1–7.
- Phillips, N. D., Hertwig, R., Kareev, Y., & Avrahami, J. (2014). Rivals in the dark: How

- competition affects information search and choices. *Cognition*, 133, 104–119.
- Rakow, T., Demes, K. A., & Newell, B. R. (2008). Biased samples not mode of presentation: Re-examining the apparent underweighting of rare events in experience-based choice. *Organizational Behavior and Human Decision Processes*, 106, 168–179.
- R Development Core Team (2008). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.
- Read, D., Loewenstein, G., & Rabin, M. (1999). Choice bracketing. *Journal of Risk and Uncertainty*, 19, 171–197.
- Redelmeier, D. A., & Tversky, A. (1992). On the framing of multiple prospects. *Psychological Science*, 3(3), 191–193.
- Samuelson, P. A. (1963). Risk and uncertainty: A fallacy of large numbers. *Scientia*, 98, 108–113.
- Stephens, D. W. (1981). The logic of risk-sensitive foraging preferences. *Animal Behavior*, 29, 628–629.
- Stephens, D. W. (2001). The adaptive value of preference for immediacy: When shortsighted rules have farsighted consequences. *Behavioral Ecology*, 12(3), 330–339.
- Stephens, D. W., Brown, J. S., & Ydenberg, R. C. (Eds.). (2007). *Foraging: Behavior and ecology*. Chicago, IL: University of Chicago Press.
- Su, Y., Rao, L. L., Sun, H. Y., Du, X. L., Li, X., & Li, S. (2013). Is making a risky choice based on a weighting and adding process? An eye-tracking investigation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 39(6), 1765–1780.
- Sutton, R. S., & Barto, A. G. (1998). *Introduction to reinforcement learning*. Cambridge, MA: MIT Press.
- Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative

- representation of uncertainty. *Journal of Risk and uncertainty*, 5(4), 297–323.
- Ungemach, C., Chater, N., & Stewart, N. (2009). Are probabilities overweighted or underweighted when rare outcomes are experienced (rarely)? *Psychological Science*, 20, 473–479.
- Unsworth, N., Heitz, R. P., Schrock, J. C., & Engle, R. W. (2005). An automated version of the operation span task. *Behavior Research Methods*, 37(3), 498–505.
- Wang, X. T., & Johnson, J. G. (2012). A tri-reference point theory of decision making under risk. *Journal of Experimental Psychology: General*, 141(4), 743–756.
- Weber, E. U., Shafir, S., & Blais, A. (2004). Predicting risk sensitivity in humans and lower animals: Risk as variance or coefficient of variation. *Psychological Review*, 111, 430–445.
- Wedell, D. H. (2011). Evaluations of single and repeated-play gambles. In J. J. Cochran (Ed.), *Wiley encyclopedia of operations research and management science*. Retrieved from <http://onlinelibrary.wiley.com/doi/10.1002/9780470400531.eorms0670/abstract>
- Wedell, D. H., & Böckenholt, U. (1990). Moderation of preference reversals in the long run. *Journal of Experimental Psychology: Human Perception and Performance*, 16, 429–438.
- Wilkinson, G. S. (1984). Reciprocal food sharing in the vampire bat. *Nature*, 308, 181–184.
- Wulff, D. U., Hills, T. T., & Hertwig, R. (2014). Online product reviews and the description–experience gap. *Journal of Behavioral Decision Making*. Advance online publication. doi:10.1002/bdm.1841

**Appendix**

Table A1.

*Additional Decision Problems*

Problem	<i>H</i>	<i>L</i>
1'	1 with $p = .75$	0 with certainty
2'	1 with certainty	0 with certainty
3'	3 with $p = .75$	9 with $p = .10$
4'	2 with certainty	7 with $p = .10$

*Note:*  $H$  = option with the higher expected value (as calculated by probability  $\times$  monetary value);  $L$  = option with the lower expected value.

## Online Product Reviews and the Description–Experience Gap

DIRK U. WULFF<sup>1\*</sup>, THOMAS T. HILLS<sup>2</sup> and RALPH HERTWIG<sup>1</sup>

<sup>1</sup>Max Planck Institute for Human Development, Berlin, Germany

<sup>2</sup>University of Warwick, Coventry, UK

### ABSTRACT

People can access information about choices in at least two ways: via summary descriptions that provide an overview of potential outcomes and their likelihood of occurrence or via sequential presentation of outcomes. Provided with the former, people make *decisions from description*; with the latter, they make *decisions from experience*. Recent investigations involving risky choices have demonstrated a robust and systematic description–experience gap. Specifically, when people make decisions from experience, rare events tend to have less impact than what they deserve according to their objective probability. Here, we show that this description–experience gap generalizes from choices involving monetary gambles to choices based on (hypothetical) online product ratings. We further show that causes that have been identified in the context of risky choice also contribute to the description–experience gap in choice based on online product ratings: reliance on relatively small samples of information and overweighting of recently sampled information (recency). We conclude with a discussion of the practical implications of our results and identify promising directions for cross-disciplinary investigations. Copyright © 2014 John Wiley & Sons, Ltd.

**KEY WORDS** word of mouth; online consumer ratings; Amazon; description–experience gap; information search; decisions under uncertainty

Neoclassical theory of consumer behavior (e.g., Marshallian demand) conceives consumer choice as choice under certainty. Challenging this conception, Savage (1954/1972) emphasized the importance of uncertainty in decisions about consumer products:

Jones is faced with the decision whether to buy a certain sedan for a thousand dollars, a certain convertible also for a thousand dollars, or to buy neither and continue carless. The simplest analysis, and the one generally assumed, is that Jones is deciding between three definite and sure enjoyments, that of the sedan, the convertible, or the thousand dollars. Chance and uncertainty are considered to have nothing to do with the situation. [...] however, it is not difficult to recognize that Jones must in fact take account of many uncertain future possibilities in actually making his choice. (pp. 83–84)

One source of uncertainty—and the one that is the concern of the present article—is the degree to which the consumer will be satisfied with his choice. Is driving a convertible really as fun as Jones expected it to be? How much will he enjoy driving it in the winter? How worried should he be about sun exposure? Fortunately, individual consumers are not alone when faced with these uncertainties. With the rise of the Internet and social media, more than ever before, consumers can learn from the experience of others. Indeed, online product reviews provide a specific form of vicarious experience that has become ubiquitous. In the fast-growing market of electronic business-to-consumer commerce (U.S. Census Bureau, 2009), they have become a market force in their own right, successfully mediating online purchasing activity (e.g., Dellarocas, 2003).

Numerous investigations have demonstrated how product reviews and ratings can affect book sales (Chevalier & Mayzlin, 2006) and box office revenues (Duan, Gu, & Whinston, 2008; Liu, 2006) or boost growth in preferences for certain types of beers (Clemons, Gao, & Hitt, 2006). To the best of our knowledge, most studies examining the link between product reviews and sales figures have analyzed large-scale panel data (e.g., Chen, Wu, & Yoon, 2004; Chevalier & Mayzlin, 2006; Dellarocas, Zhang, & Awad, 2007; Duan et al., 2008). Thus, previous research on online product reviews has predominantly taken the seller's perspective. The consumer perspective and the question of how consumers process online product ratings have received less attention. This study helps to fill this gap by taking advantage of recent findings from behavioral decision making and demonstrating how they pertain to online product reviews. It also contributes to the growing literature on online decision making (Darley, Blankson, & Luethge, 2010; Punj, 2012) that addresses, in particular, the uncertainty associated with the lack of direct experience with a product or with sales staff (Johnson, Bellman, & Lohse, 2003), as well as the information search required prior to making a selection (Horrigan, 2008; Peterson & Merino, 2003).

### Parallels between risky choice and online product reviews

Although choice between consumer products is not identical with choice between monetary gambles, there are similarities between the two: A single online consumer rating can be conceived as a potential future state of satisfaction after the purchase of the product. Thus, when a consumer seeks to buy a particular product, she may assume that her future satisfaction equals the satisfaction of the person who previously purchased the product and provided the rating. If there are many similar ratings of the product, she can assume that she will be as happy as all previous buyers.

\*Correspondence to: Dirk U. Wulff, Max Planck Institute for Human Development, Center for Adaptive Rationality (ARC), Lentzeallee 94, 14195 Berlin, Germany. E-mail: wulff@mpib-berlin.mpg.de

However, if variance occurs among raters, she will be uncertain as to which of the potential satisfaction levels will apply to her. Under the simplifying assumption that she has no additional information, she will have to assume that each individual rating in the full set of ratings (one by each rater) has the same likelihood of matching her future satisfaction level. It follows that the set of consumer ratings for a product, when aggregated by rating categories, can be understood as a gamble over states of satisfaction, where the relative frequencies of rating categories indicate the probability of future states of satisfaction.

This investigation seeks to use the resemblance between these two choice situations to create a bridge between the two fields of research. To this end, we provide an example of how the literature on risky choice can inform research on online consumer choice. Specifically, we capitalize on two dimensions that play an important role in both online product reviews and recent investigations of risky and uncertain choice: format of information presentation and distributional characteristics.

Electronic commerce (e-commerce) sites like Amazon.com display the overall mean of the available consumer ratings as a number of stars. In addition, they present a summary bar plot and a list of individual ratings. Formally, both formats present identical distributional information, but they differ substantially in the way users experience that information. Summary bar plots present complete information in one descriptive format. Individual ratings, in contrast, require the user to *sequentially search* through the ratings to acquire representative information. The distinction between summary bar plots and individual ratings can be mapped onto a distinction between two formats of information representation that has received much attention in recent investigations of risky choice involving monetary gambles. The distinction, detailed in the succeeding texts, is that between *decisions from experience* and *decisions from description* (Hertwig, Barron, Weber, & Erev, 2004). Numerous studies have demonstrated that these two kinds of formats and decisions can result in systematic and predictable differences in choices, the *description–experience gap* (for reviews, refer to Hertwig & Erev, 2009; Rakow & Newell, 2010).

The second parallel between research on risky choice and online product reviews is the *bimodal* nature of the outcome distribution. Risky choice is often studied using two-outcome gambles (Holt & Laury, 2002; Kahneman & Tversky, 1979). In many cases, these two-outcome gambles comprise a probable outcome and a complementary (relatively) rare event (Erev et al., 2010). Analyzing ratings from Amazon.com, Hu, Zhang, and Pavlou (2009) found that most distributions of online product ratings follow a J-shaped<sup>1</sup> pattern: many very positive ratings, few very negative ratings, and hardly any ratings in between. Hu et al. (2009) suggested

two selection biases to explain this distribution. First, people who give a product a low valuation are less likely to purchase it and therefore less likely to submit a rating relative to customers who actually purchased the product. Furthermore, among the purchasers, those who arrive at an extreme—either positive or negative—valuation of a product are more likely to express their views than are those with less extreme valuations, leading to a bimodal distribution (with the positive mode being more frequent than the negative one). The resulting J-shaped distribution can be conceived of as an extension of a two-outcome risky gamble containing a rare event.

In what follows, we briefly introduce relevant findings from recent research on the description–experience gap. We then explore how these findings can be brought to bear on consumer choices, based on “experienced” and “described” product reviews.

### The description–experience gap

In most studies of risky choice, people are provided with a summary description of the risky options. The options’ outcomes and associated probabilities are either conveyed visually (e.g., by a pie chart or frequency distribution) or described using numbers in text. An example of a summary description is as follows:

Option A: Receive \$4 with probability of .8, \$0 otherwise.

or

Option B: Receive \$3 for sure.

When outcomes and chances are presented in this *description* format, the majority of people choose option B (e.g., Hertwig et al., 2004; Kahneman & Tversky, 1979), even though option A has the higher expected value (A, \$3.2 vs B, \$3). This phenomenon has often been explained as a consequence of the propensity to overweight rare events; that is, people choose as if they overweight the small probability of winning nothing in gamble A (Kahneman & Tversky, 1979).

Another way to learn about the outcomes and their likelihoods is to experience those outcomes iteratively over a series of samples. For example, an onlooker witnessing the outcomes sampled from options A and B may see the following distribution of associated payoff schedules:

Option A: \$0, \$4, \$4, \$0, \$4, \$4, and \$4

Option B: \$3, \$3, \$3, \$3, \$3, \$3, and \$3

In the laboratory version of this *sampling paradigm*, participants can experience as many outcomes as they wish without the associated monetary consequences, before then deciding to terminate the exploration period and make a final choice. When gamble outcomes are presented in this *experience* format, people predominantly choose option A (Hertwig et al., 2004; Ungemach, Chater, & Stewart, 2009; but refer to Hills, Noguchi, & Gibbert, 2013). This reversal of preference implies that when decisions are based on

<sup>1</sup>The term “J-shaped” has two possible meanings: Sometimes, it is used to refer to a unimodal power-law distribution (e.g., Anderson & Schooler, 1991; Hertwig, Hoffrage, & Sparr, 2012; Todd & Gigerenzer, 2007), in which few objects have extreme values and most objects have small to medium values; sometimes, it is used to refer to a bimodal distribution (refer to Vokó et al., 1999; Witteman et al., 1994). The latter meaning is the one used here.



experience, people choose as if rare events received less weight than what they deserve in light of their objective probabilities. The description–experience gap in choice has been replicated across a wide range of studies (for reviews, refer to Hertwig & Erev, 2009; Rakow & Newell, 2010).

Why are rare events underweighted in experienced-based choices? Several reasons have been proposed (Hertwig & Erev, 2009). The two most important ones that pertain to online reviews are limited search and recency. Time constraints limit a person's ability to explore infinitely. Furthermore, there is evidence that people may be content with only small amounts of information, as small samples amplify the difference between options, thus easing choice difficulty (Hertwig & Pleskac, 2010). However, small samples also bear the risk that the decision maker is not informed about the existence of rare events or that the rare event is represented less often than expected (refer to Hertwig et al., 2004).

Another, though less powerful, factor is *recency* (compare, e.g., Hertwig et al., 2004; Rakow, Domes, & Newell, 2008; Ungemach et al., 2009). Outcomes occurring later in the sampling sequence seem to have more impact than earlier samples (Hertwig et al., 2004). This could be caused by memory limitations (e.g., Murdock, 1962) or be the outcome of an information updating process (Hogarth & Einhorn, 1992). As a consequence of recency, a decision maker who performed sampling extensively may nevertheless rely on a functionally small sample. When the functional sample size is constrained to recent samples, rare events are unlikely to be incorporated in the person's final assessment of the option.

### Does the description–experience gap generalize to choices based on online product reviews?

In summary, the situation in which people make product choices based on online product reviews has much in common with the various formats in which risky decision making has been studied. First and most importantly, in both choice situations, people make choices over probability distributions of outcomes—monetary rewards in studies on risky choice and levels of satisfaction in online consumer choice. Second, the distributions of outcomes in both situations are bimodal, with one mode being rare—usually the extreme negative mode in online product ratings. Third, the formats of information presentation used either display summary presentations of the outcomes (ratings) or require self-paced sequential search.

Despite these parallels between online consumer choice and risky choice, the two research fields have remained largely unconnected. We explore one possible link by testing whether behavioral effects documented for abstract monetary gambles generalize to choices between consumer products based on consumer ratings. The potential synergies for both domains are promising. To summarize, the rich experimental and theoretical literature on risky choice can serve as a starting point to overcome the lack of experimental work on individual consumer choice. Online consumer choice, in turn, represents an increasingly germane real-world choice scenario that can be used to test the generality of the effects found with monetary gambles. The description–experience

gap has often been demonstrated using two-outcome gambles, rendering this investigation an extension not only in terms of the type of outcome but also in terms of the pay-off distributions' complexity.

Does a description–experience gap also exist in choices based on online product ratings? To answer this question, we conducted a laboratory experiment in which participants chose between two products (e.g., camcorders) solely on the basis of product ratings. We varied the presentation of these ratings between a full summary (description format) and one requiring participants to search through a series of individual ratings (experience format). We examined the extent to which these description-based versus experience-based formats triggered systematically different choice proportions, mirroring those found in investigations of risky choice. In other words, we examined whether the experience format, relative to the description format, resulted in people choosing as if they underweighted rare (extreme) ratings relative to their objective probabilities. Moreover, we examined the extent to which two cognitive mechanisms observed as contributing to the description–experience gap in risky choice, *limited search* and *recency*, also operate in choice based on consumer ratings. Specifically, we predicted that avid searchers are more likely to have experienced rare product ratings than frugal searchers and are therefore less likely to choose as if they underweight rare ratings. In accordance with Hertwig and Pleskac's (2010) finding, we further predicted that frugal searchers will judge their decision to be easier than avid searchers, irrespective of the information experienced. Finally, we predicted that ratings experienced later in the sampling sequence will have more impact than those experienced earlier (recency effect).

## METHOD

### Participants

We collected data from 63 participants (43 female participants). The mean age was 27 years. Participants were rewarded by either course credit or a fixed payment of Confoederatio Helvetica franc (CHF) 15 (~\$15.00) and also received a monetary bonus based on the outcomes of their choices. Specifically, a random draw was taken out of each chosen rating distribution, and the resulting value of the rating (the number of stars) was multiplied by CHF 0.05. On average, participants earned a bonus of CHF 3.56 (~\$3.5).

### Procedure and material

Participants made 10 hypothetical choices between pairs of consumer products, once in the description format and once in the experience format. For each choice, product images were presented next to each other on the computer screen. We collected product images from several e-commerce sites to cover a wide range of applications and price ranges (e.g., laptops, restaurant dinners, pairs of shoes, coffee makers, etc.). The respective consumer ratings were displayed below each product (either in a summary plot or as individual ratings). Apart from consumer ratings, no further information

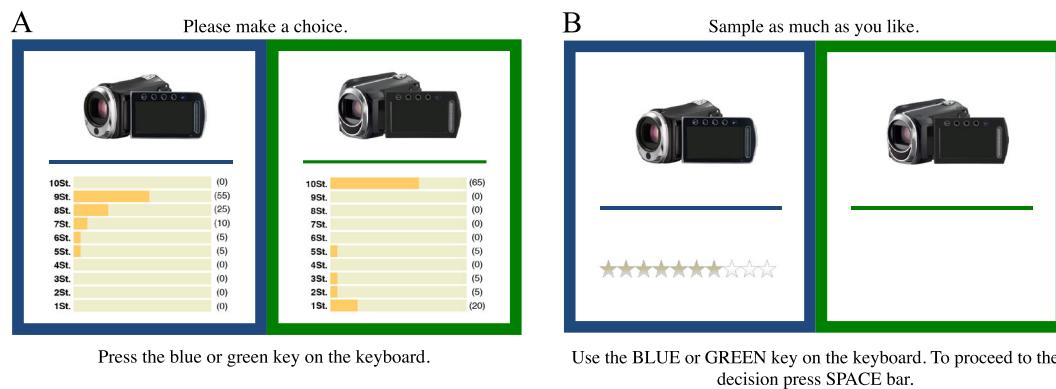


Figure 1. Screenshot of the description (A) and experience (B) rating formats. In the description format, a full table of 100 ratings is displayed, one below each product. In the experience format, ratings are presented individually below each product as it is sampled. The occurrences of individual ratings are determined by the underlying distributions (identical to those in the description format) and the search behavior of the participant.

was provided. Participants were instructed to select the product that appealed most to them given the distribution of ratings. To encourage people to pay attention to the ratings, the two pictures in each category were selected to be visually indistinguishable in terms of price, technical specifications, and quality. Participants indicated their final decision by a keyboard press.

Each participant chose twice between each pair of products—once in the *description* format and once in the *experience* format. To control for order effects, we randomized the order of the format and (right versus left) placement of products. Participants were not told that they were making the same decision twice (once in a description and once in an experience format), and the order of the two presentation formats was counterbalanced.<sup>2</sup> To further minimize the influence of prior experience, we asked participants to complete a secondary task<sup>3</sup> that took approximately 20 min between the two formats.

Figure 1 shows screenshots of the description and experience rating formats. As on the majority of e-commerce sites, ratings were displayed as stars. In the description format, the distribution of ratings was represented by a bar plot designed to resemble the summary format used on Amazon.com, in terms of color, style, and information presented (e.g., bars and counts in the description format). Each bar plot consisted of a total of 100 ratings. The total number of ratings of each star value was specified next to the bars. Participants were free to study the bar plot for as long as they wanted before making a final decision. In the experience format, participants sampled consumer ratings sequentially.

They pressed a blue or a green key to choose one or the other option and were shown a randomly sampled consumer rating for that product, displayed as a number of stars. There were no constraints in terms of time, number of samples, or sampling sequence. The ratings were randomly drawn with replacement from the underlying hidden distribution of ratings, which was identical to that presented in the description format. Participants indicated when they were ready to stop sampling. Once sampling was terminated, they were asked to make their final choice.

Figure 2 displays an example pair of the distributions employed. In every pair of options, one was clearly unimodal (A). The other option (B) was bimodal and followed the J-shaped pattern described in Hu et al. (2009). These distributions allowed us to study the psychological impact of rare ratings (refer to APPENDIX B for a full table of the choice problems used). For example, based on the complete distribution of ratings in Figure 2, option A has the higher mean rating. However, assuming the rare ratings at the most negative end of option B has little or no impact—because they are not sampled, undersampled, or not recently sampled—then, option B will have the higher experienced mean and may thus be preferred over option A. For all pairs of distributions, it holds that *not* choosing the higher objective mean (HOM) is consistent with underweighting rare product ratings. Put differently, one option always represented the (objectively) higher mean rating; the other option represented the (objectively) higher median rating.

In addition to sampling and choice data, we collected information on the perceived difficulty of a choice. Specifically, participants rated the difficulty of each choice on a scale from 1 (*very easy*) to 5 (*very difficult*).

## RESULTS

Two of the 10 distribution pairs were incorrectly specified in our automated protocol for a substantial part of the data collection. The following results are therefore based on only eight of the 10 product choices per format.

<sup>2</sup>The order in which participants worked through the two formats did not affect either choice proportions (description,  $t_{61} = 0.93$ ;  $p = .355$ ; experience,  $t_{61} = 1.34$ ;  $p = .185$ ) or average sample sizes ( $t_{61} = 1.52$ ,  $p = .133$ ).

<sup>3</sup>The secondary task was the automated operation span task developed by Unsworth, Heitz, Schrock, and Engle (2005). We chose this task for two reasons. First, it is a rather long task (~20 min), making carry-over effects from one format to the other relatively unlikely. Second, one previous study has reported a relationship between working memory capacity and sample size (Rakow et al., 2008). We investigated whether this finding could be replicated using a similar working memory task. However, we found no relationships between operation span, as follows: (i) sample size ( $r = .04$ ); (ii) switch rate ( $r = -.07$ , refer to Hills & Hertwig, 2010); or (iii) subsample size ( $r = -.02$ ).



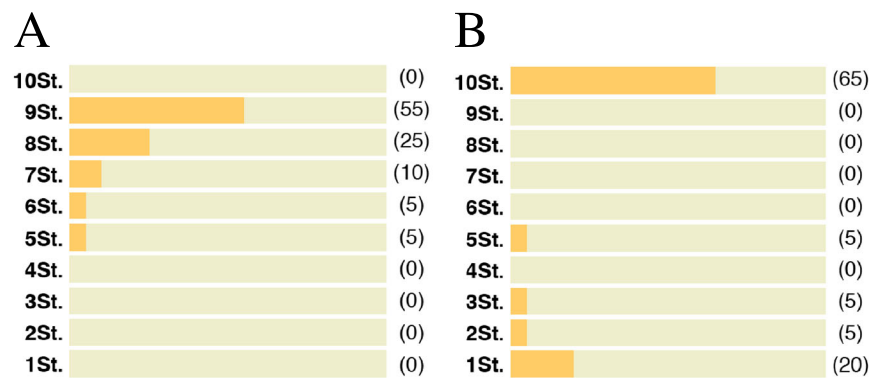


Figure 2. Pair of distributions of consumer ratings. In this example, Distribution A is superior in terms of the mean star rating and therefore more likely to be chosen in the absence of underweighting of rare events. Underweighting rare events, in contrast, should result in favoring Option B (J-shaped distribution).

### Is there a description–experience gap in choice based on consumer ratings?

The description and experience rating formats resulted in substantially different choices in relation to the products' objective mean rating. Figure 3 shows that the probability of choosing the product with the HOM rating was about 13 percentage points lower when the choice was based on experience ( $M=65.5\%$ ,  $SD=19.3\%$ ) as opposed to description ( $M=78.4\%$ ,  $SD=24\%$ ). Thus, even though participants saw the same product options in the experience and description formats, which were both based on the same underlying distributions, the participants chose the HOM option less often when their decisions were based on the experience format,  $t(62)=3.66$ ,  $p<.001$ ,  $d=.59$ .<sup>4</sup>

This behavior is consistent with people in experience-based risky choice choosing as if rare events receive less weight than what they deserve according to their objective probability (Hertwig & Erev, 2009; Hertwig et al., 2004). The description–experience gap thus appears to generalize from the domain of monetary gambles to the domain of online consumer choice based on product ratings. Next, we examine to what extent the gap can be explained in terms of small samples and recency.

### Small samples

Probably, the most important factor in the gap between the description and experience formats is limited search in the experience format (Hertwig & Erev, 2009). Small samples reduce the likelihood of encountering rare ratings (be they positive or negative) and thereby reduce their impact.<sup>5</sup> The

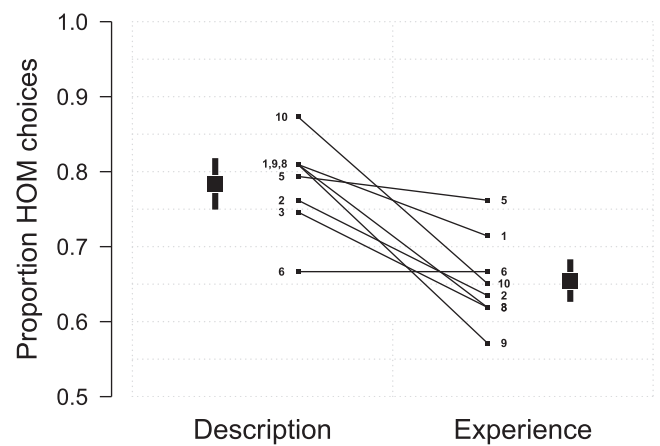


Figure 3. Description–experience gap. Overall proportions of people choosing the higher objective mean separately for description and experience format of consumer ratings are displayed. Lines and numbers represent the decision proportions for the eight problems analyzed. Error bars represent standard error of the mean.

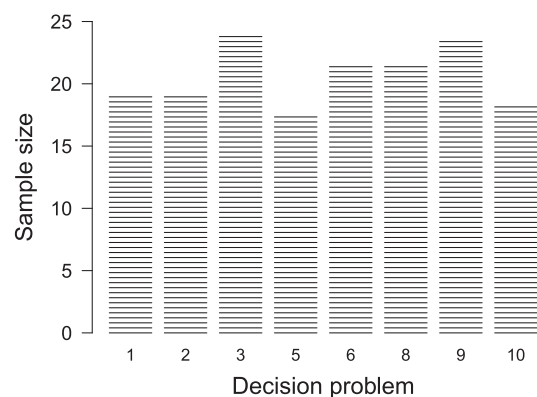


Figure 4. Sample size per decision problem.

average sample size per decision problem varied between 18 and 24 (Figure 4), with a mean of 20.67 ( $SD=2.46$ ). These numbers are similar to but slightly larger than the sample sizes reported in other computer-based studies of decisions from experience (e.g., Hertwig & Pleskac, 2010; Hertwig et al., 2004). One possible reason for this small increase in sample size is the absence of sure options, which are usually explored less extensively (Lejarraga, Hertwig,

<sup>4</sup>Mixed-effects analyses revealed that the inclusion of a fixed problem factor did not improve the prediction of choices in either the experience or the description format (likelihood-ratio test; experience,  $X^2_7=7.45$ ,  $p=.38$ ; description,  $X^2_7=12.63$ ,  $p=.08$ ).

<sup>5</sup>The likelihood of experiencing a rating can be understood in terms of the proportion of people that observe the rating less frequently than expected or not at all. Thus defined, a reduced likelihood can be qualified via the shape of the sampling distribution: A right-skewed sampling distribution implies a higher proportion of people experiencing a rating less often than the expected value, and vice versa. The sampling distribution for the number of occurrences of any outcome is governed by a binomial distribution and its skewness is calculated as  $\frac{1-2p}{\sqrt{n^*p^*(1-p)}}$ . This term is positive (right skewed) for all  $p<.5$  and increases with smaller  $p$ s and smaller  $n$ s. Hence, smaller sample sizes reduce the likelihood of experiencing rare ratings.

& Gonzalez, 2012). Critically, the observed sample sizes are sufficiently small to render it possible for small sample size to have a direct impact on choice. For illustration, with 20 draws spread across both products, the chances of experiencing each of five possible star ratings in both options (the maximum per option in this study), assuming an equal distribution of ratings (each rating has a likelihood of 20%), are about one in five. Thus, small sample sizes could easily influence choice.

Indeed, small samples had a direct impact on the final choice. Taking more samples increased the likelihood that participants would choose the option with the HOM rating (the correlation between mean individual sample size and the proportion of HOM mean choices was  $r = .47$ ,  $p < .001$ ). A mediation analysis showed that the reduced sampling of rare ratings was sufficient to explain the fewer choices favoring the HOM ratings (APPENDIX A). Thus, consistent with our predictions, one explanation for the description–experience gap is that participants were content with relatively small samples of ratings in the experience format, thus either missing the rare ratings or experiencing them less frequently than expected. This led participants in experience-based formats to make choices as if they were underweighting rare ratings.

Why do people content themselves with relatively small samples of information that is essentially free? One possible explanation is the *amplification effect* (Hertwig & Pleskac, 2010): Small samples amplify the perceived difference between the expected mean earnings associated with the payoff distributions, thus making the options more distinct and choice easier. The same argument applies to distributions of consumer ratings. Consistent with the amplification effect, we found that our participants' evaluations of choice difficulty were substantially correlated with sample size (subject level:  $r = .39$ ,  $p = .002$ ). Specifically, avid searchers judged decisions to be more difficult than did frugal searchers. However, a within-participant comparison between the experience and description formats revealed that choice difficulty for frugal searchers was not attenuated as has been observed by Hertwig and Pleskac (2010). Following the original Hertwig and Pleskac analysis, we analyzed perceived difficulty as a function of a median split on sample size and the different formats. Frugal searchers judged their experience-based choices to be as easy as those made in the description format,  $t(30) = 1.59$ ,  $p = .12$ . Avid searchers, in contrast, judged their experience-based choices to be significantly more difficult than those made in the description format,  $t(31) = 4.57$ ,  $p < .001$ . However, the interaction failed to reach significance,  $F(1,120) = 0.4$ ,  $p = .85$ . Overall, drawing higher numbers of samples nevertheless appeared to be associated with decreased ease of making decisions.

Of course, this analysis ignores some important information. It glosses over the stratified nature of the data and neglects the mediating role of actual difficulty (i.e., the difference in experienced means between problems). If the amplification effect works as proposed, then higher levels of search should be associated with increased perceived difficulty, and this association should, in turn, be mediated by the actual difficulty. Alternatively, if difficulty is a mere expression of effort, then taking more or fewer samples should remain related to perceived

difficulty even after the inclusion of actual difficulty. To address this issue, we performed a mixed-effects analysis<sup>6</sup> predicting the perceived difficulty by sample size and, in the second step, the final Cohen's  $d$  based on the experienced outcomes of a given problem as an indicator of actual difficulty. To account for dependent measurements, we included two random intercepts for participants and problems.

Consistent with the amplification effect, we found sample size to be highly associated with perceived difficulty ( $b = .57$ ,  $p < .001$ ). Importantly, this association was only moderately reduced by the inclusion of actual difficulty (partial effect:  $b = .49$ ,  $p < .001$ ). Thus, the effort of sampling appears to contribute to the perception of difficulty. However, the pattern of partial mediation is completed by significant associations both between sample size and actual difficulty ( $b = -.29$ ,  $p < .001$ ), with larger sample sizes being related to smaller differences, and between actual difficulty and perceived difficulty ( $b = -.16$ ,  $p < .001$ ; Baron & Kenny, 1986; refer also to APPENDIX A). Thus, in addition to the influence of effort, small samples rendered choices easy.

### The influence of recency on consumer choice

In past research, recency has not consistently been observed to affect decisions from experience (refer to Hertwig & Erev, 2009). Did it affect our participants' product choices? We based our analysis on the initial and final samples taken from each option. Sample means were computed for both options' initial and final sampling periods (Figure 5) and compared with respect to their ability to predict the final choices.

Out of 96 cases where the initial and final sampling period suggested different options to be better, participants chose the option that had the higher mean in the most recent sampling period in 72 (74%) cases (sign test,  $p < 0.001$ ). Moreover, a mixed-effects analysis using the means within the *first samples* and *last samples* also revealed a much higher impact for the mean difference in the last samples (*odds ratio* = 17.27,  $p < .001$ ) than that for the mean difference in the first samples (*odds ratio* = 2.31,  $p < .001$ ).<sup>7</sup> Recency thus appears to have played an important role in product choice based on online consumer ratings.

## DISCUSSION

Social information in the form of consumer ratings is a driving factor behind online consumer choice. We

<sup>6</sup>Mixed-effects analyses were performed using the R packages *lme4* and *lmerTest*. Degrees of freedom for the Gaussian linear models were estimated using Satterthwaite's approximation, the default method in *lmerTest*. For better comparison, all predictors were standardized.

<sup>7</sup>This analysis was based on 296 of 504 decisions in which the participants switched at least twice between the options. We also tested whether this effect was moderated by differences in the number of samples in the *first samples* (average length = 12.4) and *last samples* (average length = 9). However, the inclusion of two variables representing the number of samples left the effects unchanged and did not result in improved model fit,  $X^2(2) = 4.44$ ,  $p = .11$ .

Sample	1	2	3	4	5	6	7	8	9	10	11
Option 1	10	10	5					2		10	10
Option 2				8	9	9	9			9	

*First samples*
*Last samples*

Figure 5. Illustration of *first samples* and *last samples*. *First samples* include all samples prior to the second switch, and *last samples* include all samples beyond the second to last switch.

investigated the extent to which recent findings in research on decisions from experience in the domain of monetary gambles generalize to choice based on online consumer ratings. Our results suggest that the domain of online consumer choice may be subject to some of the same information-format dependence as observed in risky choice. There is a profound difference between making choices based on a summary “descriptive” format of online consumer ratings and making choices based on sequential sampling from individual consumer ratings, even when the underlying distributions of the ratings are the same (Hertwig & Erev, 2009).

Our results further demonstrate that factors previously proposed to contribute to the description–experience gap may apply more generally. Specifically, we observed three contributors to the description–experience gap in choice based on online consumer ratings: First, people perceived choices to be easier when they took smaller samples (refer to Hertwig & Pleskac, 2010). Second, small sample sizes reduced the likelihood of participants experiencing rare information, leading them to make choices as if they underweighted rare ratings. Third, participants were clearly influenced by the recency of sampled information (Hertwig et al., 2004)—again, leading them to make choices as if they underweighted rare ratings. In sum, the full set of core findings on the description–experience gap persisted in a (hypothetical) online consumer choice scenario in which the outcome distributions were more complex than in previous investigations of the description–experience gap. This not only opens up many new directions for future research but also has specific implications for e-commerce.

In particular, the format dependence of the impact of infrequent ratings is of great importance for e-commerce. As noted by Hu et al. (2009), the majority of consumer rating distributions are J-shaped, with many favorable ratings and few unfavorable ones. Our findings indicate that this will lead people to have lower expectations of consumer goods when looking at summary description-based formats than when perusing individual ratings or entries (but refer to Ert, 2005). Administrators of e-commerce sites can potentially use these findings to foster more informed consumer choice and consumer satisfaction by making sure that consumers always have access to full summary descriptions. Further, the observed recency effect illustrates the relevance of presentation order of consumer ratings. Finally, our findings are relevant for the growing problem of separating truthful from fabricated reviews (Streitfeld, 2013). If fake ratings are both extreme *and* rare, then the use of the experience format would

naturally undermine their influence in much the same way as a trimmed mean reduces the influence of strategic scoring in sports competitions (refer to Bamberger, Erev, Kimmel, & Oref-Chen, 2005).

Further, there is a rich set of findings in research on decisions from experience involving risky choice that appears relevant to research on the psychological impact of online product reviews. For instance, it has been demonstrated that the amount of information search substantially varies with factors such as the decision maker's affective state (Frey, Hertwig, & Rieskamp, 2014), the value of the options (Hau, Pleskac, Kiefer, & Hertwig, 2008), the choice domain (i.e., gain versus loss; Lejarraga et al., 2012), and the influence of prior sampling from larger or smaller set sizes (Hills et al., 2013). Another important finding is that the way people search for information in terms of switching between options (or distribution of ratings) foreshadows the decision strategies that people appear to use (Hills & Hertwig, 2010)—providing another potential explanation for the description–experience gap. Specifically, it has been shown that people who often switch between options in the sampling period do not maximize the mean outcome but rather tend to choose an option that is better “most of the time.” Finally, in light of the inconsistency of previous findings on recency effects (e.g., Rakow et al., 2008; Ungemach et al., 2009), the pronounced recency effect observed here suggests problem complexity (e.g., number of distinct outcomes/ratings) as a potential moderator of recency in experience-based choice.

Of course, we should emphasize that this first investigation does not reflect the true complexity of e-commerce sites. Most importantly, the majority of sites (e.g., Amazon.com, Tripadvisor.com, etc.) allow consumers to peruse ratings in combination with written reviews. These range from largely uninformative brief statements (“great book”) to reviews providing valuable assessments of a product and its properties. Our investigation cannot account for this or for other sources of information (e.g., ratings of the helpfulness of a review, full profiles of reviewers, and total number of ratings). All of these dimensions can and should be addressed in subsequent studies.

Last but not least, we should emphasize that our study—based on hypothetical product reviews and incentivized, but ultimately hypothetical, choices between pairs of consumer products—cannot approximate the rich motivational structure of actual consumer choice. The goals of people buying consumer products of the type used here (e.g., camcorders) will differ from those of our participants. First, a consumer may focus on a single product (rather than two or more

products) and is likely to compare products along numerous potentially incommensurable dimensions. Second, as a consumer typically purchases only one, say, camcorder, he or she may aim to minimize the maximum loss (the purchase of a “lemon”) or to satisfy an aspiration level for each purchase. In contrast, in our study implementing 10 choices, a bad outcome in one choice can be compensated by a good outcome in another; hence, the participant can aggregate the risk over choices bracketed together (Read, Loewenstein, & Rabin, 1999). Therefore, the robustness of the present results should next be tested in settings with real product ratings, real consumer products, and real choices. Notwithstanding these issues, however, it is worth noting that the description–experience gap obtained in monetary gambles, and replicated here, has also been found in (hypothetical) choices in which people relied on a minimax heuristic (thus avoiding the worst possible outcome), namely, in choices between drugs with different uncertain side effects (Lejarraga, Pachur, Frey, & Hertwig, 2014). Furthermore, individual choice problems in a collection of problems are often played as if they were faced in isolation (Wulff, Hills, & Hertwig, 2014). These results raise the possibility that the gap between choices in the laboratory and consumer choices is perhaps smaller than it might first appear.

The aim of this article has been to relate two hitherto unrelated lines of research on human choice, namely, online consumer choice and risky choice between monetary gambles. The literature on risky choice has produced a large body of experimental findings and theoretical explanations. We found that some key findings on the description–experience gap in risky choice generalize to online consumer choice. This raises the promising and fruitful possibility that other effects observed in research on experience-based and description-based risky choice may also generalize to consumer choice. If so, human choice across different domains may, to some extent, follow the same regularities.

APPENDIX A

The impact of sample size on higher objective mean (HOM) choices was predicted to be mediated by experiencing versus not experiencing rare events. To test this prediction, we performed a mediation analysis on the trial level with the percentage of possible distinct ratings experienced as the mediator. Specifically, we specified a mixed-effects model via the lmer and glmer functions in the R package lme4, with random subject intercepts and standardized variables. We found that sample size significantly predicted HOM (*odds ratio* = 1.32, *p* = .011) and the percentage of distinct ratings experienced ( $\beta$  = .53, *p* < .001), with higher sample sizes leading to more HOM and the observation of more distinct ratings (Figure A1). Thus, two of the necessary conditions in Baron and Kenny’s steps for mediation (Baron & Kenny, 1986) are fulfilled. The third condition postulates that the size of the direct effect of the independent variable (sample size) on the dependent variable (HOM) either drops substantially after the inclusion of the mediator (partial mediation) or vanishes completely (full mediation). We found the latter. The effect of sample size on HOM vanished entirely (*odds ratio* = 1.02, *p* = .865) when we controlled for the percentage of distinct ratings experienced. Thus, the effect of sample size on HOM was fully mediated by the percentage of distinct ratings experienced.

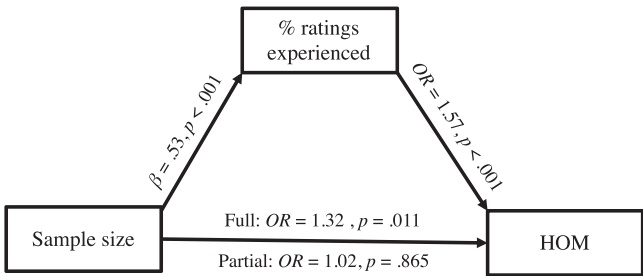


Figure A1. Mediation analysis. “Full” indicates the effect of sample size on higher objective mean (HOM) choices without controlling for percentage of distinct ratings experienced, whereas “Partial” indicates the effect when the mediator is accounted for.

APPENDIX B

Table B1. Choice problems (P) employed in our investigation

Stars	P1		P2		P3		P4		P5		P6		P7		P8		P9		P10	
	L	H	L	H	L	H	L	H	L	H	L	H	L	H	L	H	L	H	L	H
1	.2					.65			.35		.05	.05	.25		.1		.6		.3	
2	.05				.9		.25				.65	.05			.1		.6	.05		.05
3	.05			.7	.05		.05				.7		.05		.6		.35	.05		
4			.9			.05					.15			.05	.8					
5	.05	.05	.1				.1				.1			.05						.05
6		.05					.05		.15				.1			.05				.05
7		.1			.1				.85				.75							
8		.25							.65				.55	.05						.05
9		.55			.15	.05	.8					.1			.05		.15			.85
10	.65		.3	.05	.05	.65	.05				.3				.25		.15	.65		

Note: Entries to left and right of the shaded lines denote the relative frequencies/probabilities of the 1 to 10 star values. Problems 3 and 7 were miscoded for the first 13 of the 63 participants; the numbers displayed correspond to the problems seen by the remaining 50 participants. In order to remedy this mistake, we restricted the analyses to the other eight problems.



## ACKNOWLEDGEMENTS

We thank Yvonne Bennett and Susannah Goss for editing the manuscript and the Swiss National Science Foundation for a grant to the second author (100014\_130397) and the third author (100014–126558).

## REFERENCES

- Anderson, J. R., & Schooler, L. J. (1991). Reflections of the environment in memory. *Psychological Science*, 2, 396–408.
- Bamberger, P. A., Erev, I., Kimmel, M., & Oref-Chen, T. (2005). Peer assessment, individual performance, and contribution to group processes: The impact of rater anonymity. *Group & Organizational Management*, 30, 344–377.
- Baron, R. M., & Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, 51(6), 1173–1182.
- Chen, P., Wu, S., & Yoon, J. (2004). The impact of online recommendations and consumer feedback on sales. *ICIS 2004 Proceedings*, 711–723.
- Chevalier, J. A., & Mayzlin, D. (2006). The effect of word of mouth on sales: Online book reviews. *Journal of Marketing Research*, 43, 345–354.
- Clemons, E. K., Gao, G., & Hitt, L. M. (2006). When online reviews meet hyperdifferentiation: A study of the craft beer industry. *Journal of Management Information Systems*, 23, 149–171.
- Darley, W. K., Blankson, C., & Luethge, D. J. (2010). Toward an integrated framework for online consumer behavior and decision making process: A review. *Psychology & Marketing*, 27, 94–116.
- Dellarocas, C. (2003). The digitization of word of mouth: Promise and challenges of online feedback mechanisms. *Management Science*, 49, 1407–1424.
- Dellarocas, C., Zhang, X., & Awad, N. F. (2007). Exploring the value of online product reviews in forecasting sales: The case of motion pictures. *Journal of Interactive Marketing*, 21, 23–45.
- Duan, W., Gu, B., & Whinston, A. B. (2008). Do online reviews matter? An empirical investigation of panel data. *Decision Support Systems*, 45, 1007–1016.
- Erev, I., Ert, E., Roth, A. E., Haruvy, E., Herzog, S. M., Hau, R., Hertwig, R., Stewart, T., West, R., & Lebiere, C. (2010). A choice prediction competition: Choices from experience and from description. *Journal of Behavioral Decision Making*, 23, 15–47.
- Ert, E. (2005). Seeing is believing: The positive and negative effects of free sampling. *Unpublished Master's Thesis*, Technion, Israel.
- Frey, R., Hertwig, R., & Rieskamp, J. (2014). Fear shapes information acquisition in decisions from experience. *Cognition*, 132(1), 90–99.
- Hau, R., Pleskac, T. J., Kiefer, J., & Hertwig, R. (2008). The description–experience gap in risky choice: The role of sample size and experienced probabilities. *Journal of Behavioral Decision Making*, 21, 493–518.
- Hertwig, R., Barron, G., Weber, E. U., & Erev, I. (2004). Decisions from experience and the effect of rare events in risky choice. *Psychological Science*, 15, 534–539.
- Hertwig, R., & Erev, I. (2009). The description–experience gap in risky choice. *Trends in Cognitive Science*, 13, 517–523.
- Hertwig, R., Hoffrage, U., & Sparr, R. (2012). How estimation can benefit from an imbalanced world. In P. M. Todd, G. Gigerenzer, & the ABC Research Group (Eds.), *Ecological rationality: Intelligence in the world* (pp. 379–409). New York, NY: Oxford University Press.
- Hertwig, R., & Pleskac, T. J. (2010). Decisions from experience: Why small samples? *Cognition*, 115, 225–237.
- Hills, T. T., & Hertwig, R. (2010). Information search in decisions from experience: Do our patterns of sampling foreshadow our decisions? *Psychological Science*, 21, 1787–1792.
- Hills, T. T., Noguchi, T., & Gibbert, M. (2013). Information overload or search-amplified risk? Set size and order effects on decisions from experience. *Psychonomic Bulletin & Review*, 20, 1023–1031.
- Hogarth, R. M., & Einhorn, H. J. (1992). Order effects in belief updating: The belief-adjustment model. *Cognitive Psychology*, 24, 1–55.
- Holt, C. A., & Laury, S. K. (2002). Risk aversion and incentive effects. *The American Economic Review*, 92, 1644–1655.
- Horrigan, J. B. (2008). *Online shopping*. Washington, DC: Pew Internet Life & American Project Report.
- Hu, N., Zhang, J., & Pavlou, P. A. (2009). Overcoming the J-shaped distribution of product reviews. *Communications of the ACM*, 52, 144–147.
- Johnson, E. J., Bellman, S., & Lohse, G. L. (2003). Cognitive lock-in and the power law of practice. *Journal of Marketing*, 67, 62–75.
- Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, 47, 263–292.
- Lejarraga, T., Hertwig, R., & Gonzalez, C. (2012). How choice ecology influences search in decisions from experience. *Cognition*, 124, 334–342.
- Lejarraga, T., Pachur, T., Frey, R., & Hertwig, R. (2014). Decisions from experience: From monetary to medical gambles. Manuscript submitted for publication.
- Liu, Y. (2006). Word-of-mouth for movies: Its dynamics and impact on box office revenue. *Journal of Marketing*, 70, 74–89.
- Murdock, B. B. (1962). The serial position effect of free recall. *Journal of Experimental Psychology*, 64, 482–488.
- Peterson, R. A., & Merino, M. C. (2003). Consumer information search behavior and the Internet. *Psychology & Marketing*, 20, 99–121.
- Punj, G. (2012). Consumer decision making on the web: A theoretical analysis and research guidelines. *Psychology & Marketing*, 29, 791–803.
- Rakow, T., Demes, K. A., & Newell, B. R. (2008). Biased samples not mode of presentation: Re-examining the apparent underweighting of rare events in experienced-based choice. *Organizational Behavior and Human Processes*, 106, 168–179.
- Rakow, T., & Newell, B. R. (2010). Degrees of uncertainty: An overview and framework for future research on experience-based choice. *Journal of Behavioral Decision Making*, 14, 1–14.
- Read, D., Loewenstein, G., & Rabin, M. (1999). Choice bracketing. *Journal of Risk and Uncertainty*, 19, 171–197.
- Savage, L. J. (1954/1972). *The foundations of statistics*. New York, NY: Dover.
- Streitfeld, D. (2013, September 22). Give yourself 5 stars? Online, it might cost you. *The New York Times*. Retrieved from [http://www.nytimes.com/2013/09/23/technology/give-yourself-4-stars-online-it-might-cost-you.html?\\_r=0](http://www.nytimes.com/2013/09/23/technology/give-yourself-4-stars-online-it-might-cost-you.html?_r=0)
- Todd, P. M., & Gigerenzer, G. (2007). Environments that make us smart: Ecological rationality. *Current Directions in Psychological Science*, 16(3), 167–171.
- U.S. Census Bureau. (2009). Annual retail trade report. Retrieved from <http://www.census.gov/retail>
- Ungemach, C., Chater, N., & Stewart, N. (2009). Are probabilities overweighted or underweighted when rare outcomes are experienced (rarely)? *Psychological Science*, 20, 473–479.
- Unsworth, N., Heitz, R. P., Schrock, J. C., & Engle, R. W. (2005). An automated version of the operation span task. *Behavior Research Methods*, 37, 498–505.
- Vokó, Z., Bots, M. L., Hofman, A., Koudstaal, P. J., Witteman, J. C. M., & Breteler, M. M. B. (1999). J-shaped relations between blood pressure and stroke in treated hypertensives. *Hypertension*, 34, 1181–1185.
- Witteman, J. C. W., Grobbee, D. E., Volakenburg, H. A., van Hemert, A. M., Stijnen, T., Burger, H., & Hofman, A. (1994). J-shaped relation between change in diastolic blood pressure and progression of aortic atherosclerosis. *The Lancet*, 343, 504–507.
- Wulff, D., Hills, T. T., & Hertwig, R. (2014). The impact of short- and long-run frames on search and choice in decisions from experience. Manuscript submitted for publication.

*Authors' biographies:*

**Dirk U. Wulff** is a Predoctoral Fellow at the Center for Adaptive Rationality (ARC) at the Max Planck Institute for Human Development in Berlin, Germany. He received his Masters level degree from the University of Marburg, Germany, focusing on cognitive psychology, clinical psychology and statistics.

**Thomas T. Hills** is a Professor at the Department of Psychology at the University of Warwick in Coventry, UK. He received his PhD in Biology from the University of Utah. His research focuses on information search, information structure, and their combined influence on learning and memory.

**Ralph Hertwig** is the Director of the ARC at the Max Planck Institute for Human Development in Berlin, Germany. He held

positions at the University of Chicago and Columbia University and served as the chair for Cognitive and Decision Science and the Dean of Research at the University of Basel.

*Authors' addresses:*

**Dirk U. Wulff**, Max Planck Institute for Human Development, Berlin, Germany.

**Thomas T. Hills**, University of Warwick, Coventry, UK.

**Ralph Hertwig**, Max Planck Institute for Human Development, Berlin, Germany.



Prof. Dr. Ralph Hertwig ist Direktor des Forschungsbereichs Adaptive Rationalität am Max-Planck-Institut für Bildungsforschung.



Dipl.-Psych. Dirk U. Wulff ist Doktorand am Forschungsbereich Adaptive Rationalität am Max-Planck-Institut für Bildungsforschung.

# Risikoentscheidungen: Die Kluft zwischen Erfahrung und Beschreibung

Ralph Hertwig und Dirk U. Wulff

Dieser Artikel gibt eine Einführung in das „Description-Experience Gap“ – der Beobachtung eines systematischen Unterschieds in Risikoentscheidungen aufgrund symbolischer Beschreibungen bzw. sequentieller Erfahrungen. Die „Kluft“ wird besonders deutlich bei seltenen Ereignissen, die zu viel Gewicht (Beschreibung) bzw. zu wenig Gewicht (Erfahrung) erhalten können. Ursachen und Implikationen werden diskutiert.

## 1. Das „Gewicht“ seltener Ereignisse

In den Jahren nach 2007 rang die Welt mit der Gefahr einer Kernschmelze des internationalen Finanzsystems. Mehrere extreme Ereignisse, von denen jedes für sich genommen sehr unwahrscheinlich schien – der Zusammenbruch der U.S. Bank *Lehman Brothers*, der drohende Bankrott großer Banken (z. B. *Hypo Real Estate*), großer Finanzdienstleister (z. B. *AIG*, *Fannie Mae*, *Freddie Mac*) und ganzer Staaten –, folgten in schneller Abfolge aufeinander. Was als eine Krise des internationalen Finanzsystems begann, wurde schnell zu einer Weltwirtschaftskrise, in deren Folge globale Unternehmen in Schieflage gerieten, Staatsbudgets aus dem Ruder liefen und die Arbeitslosigkeit in vielen Ländern rasant anstieg. Warum war die Welt auf die Möglichkeit dieser Ereignisse so schlecht vorbereitet? Eine Erklärung lautet, dass die Risikomanagementmodelle mit ihrer auf der Normalverteilungshypothese aufbauenden Logik in jenen Welten scheitern, die systematisch von der Gauß-Funktion abweichen (z. B. „fat tails“-Verteilungen). In Welten, in denen selten doch nicht so selten ist wie theoretisch angenommen, unterschätzen diese Modelle die Häufigkeit seltener, extremer Verluste (vgl. *Taleb*, 2007).

Das so angeblich optimierte Risikomanagement ist aber sicher nicht das alleinige Glied in der Ursachenkette. Auch die individuellen Marktteilnehmer haben sich offensichtlich nicht genügend auf das Risiko extremer ökonomischer Ereignisse eingestellt – zum Beispiel all jene

Besitzer einer Immobilie, die im Verlauf der Krise ihre Kredite nicht mehr bedienen konnten und sich mit einer Zwangsversteigerung konfrontiert sahen. Warum aber haben viele Akteure durchweg die Möglichkeit seltener, extremer Ereignisse unterschätzt oder gar völlig ignoriert? Können psychologische Theorien und Befunde – letztere gewonnen in Verhaltensexperimenten zu Entscheidungen unter Risiko und Unsicherheit – diese mutmaßliche Achtlosigkeit gegenüber seltenen, aber schwerwiegenden Ereignissen erklären? Auf den ersten Blick nicht. Zahlreiche Studien der Entscheidungsforschung scheinen eher eine gegenteilige Tendenz zu belegen: Die Wahrscheinlichkeit seltener Ereignisse wird häufig überschätzt. Zum Beispiel werden relativ seltene Risiken, wie die einer Lebensmittelvergiftung oder Lungenkrebs infolge von Nikotingenuss, im Schnitt viel zu hoch eingeschätzt (vgl. *Lichtenstein et al.*, 1978; *Viscusi*, 2002). Mediale Berichterstattung, die dazu neigt, seltenen, aber dramatischen Krankheiten und Todesursachen (z. B. BSE und die Creutzfeld-Jakob-Krankheit) unverhältnismäßig viel Beachtung zu schenken, spielt bei dieser Tendenz zur Überschätzung sicher eine Rolle (vgl. *Renn*, 2014). Aber selbst wenn seltene Risiken nicht geschätzt werden müssen, sondern explizit quantifiziert sind, wird unwahrscheinlichen Ereignissen mehr psychologisches Gewicht beigemessen als ihnen, gemessen an ihrer objektiven Wahrscheinlichkeit, zustünde. Dies ist zumindest eine der zentralen Annahmen in der *Cumulative Prospect-Theorie* (vgl. *Tversky/Kahneman*, 1992), der einflussreichsten deskriptiven Theorie des Entscheidens unter Risiko. Mithil-

### Stichwörter

- Entscheidungen unter Risiko und Unsicherheit
- Erfahrungs- und beschreibungs-basierte Entscheidungen
- Exploration
- Psychologie seltener Ereignisse
- Risikokommunikation

fe der Annahme einer nichtlinearen Wahrscheinlichkeitsgewichtungsfunktion, die einen umgekehrt S-förmigen Verlauf nimmt und nach der seltene Ereignisse „übergewichtet“ und Ereignisse mit mittleren und hohen Wahrscheinlichkeiten „untergewichtet“ werden, erklärt sie Verhaltensanomalien, die die **Erwartungsnutzentheorie** vor Probleme stellt (z. B. gleichzeitige Risikofreude und Risikoaversion in Gestalt von Lottospielen und Erwerb von Versicherungen).

Warum also handeln Menschen, die nach weithin akzeptierter Sichtweise dazu neigen, geringe Wahrscheinlichkeiten zu überschätzen oder diesen, sobald sie expliziert werden, zu viel Gewicht einzuräumen, als ob sie die Möglichkeit seltener katastrophaler Ereignisse nicht ernst genug nähmen? Um dieses scheinbare Paradox zu verstehen, hilft eine Untersuchung, die seit etwa einer Dekade in der Entscheidungsforschung zunehmend Beachtung findet: Über die Wahrscheinlichkeit eines unsicheren Ereignisses kann man – sehr vereinfacht – auf zwei grundsätzlich unterschiedlichen Wegen Kenntnis erlangen: durch die symbolische **Beschreibung** seiner Wahrscheinlichkeit oder durch die sequentielle **Erfahrung** des Auftretens bzw. Nichtauftretens des Ereignisses (vgl. Hertwig et al., 2004). Die Annahme der Überbewertung seltener Ereignisse lässt außer Acht, dass eine Vielzahl von Entscheidungen nicht auf expliziten und beschriebenen Informationen über Wahrscheinlichkeiten beruhen. In diesen Fällen, die nicht die Ausnahme, sondern eher die Regel darstellen, können Menschen häufig nichts anderes tun, als sich auf ihre Erfahrungen zu verlassen.

Ziel dieses Beitrages ist es, Erkenntnisse der psychologischen Forschung zur „Kluft“ zwischen erfahrungs- und beschreibungsbasierten Risikoentscheidungen zu erläutern und Konsequenzen für das unternehmerische Risikomanagement zu diskutieren. Im nächsten Abschnitt beschreiben wir hierzu die Grundlagen erfahrungs- und beschreibungsbasierter Risikoentscheidungen. In Abschnitt 3 stellen wir die experimentelle Methodik vor, die verwendet wird, um diese systematische Kluft zwischen den beiden Formaten zu untersuchen. Anschließend umreißen wir in Abschnitt 4 zwei wesentliche Erklärungen für den Unterschied zwischen den beiden Klassen

von Entscheidungen. Zum Abschluss veranschaulichen wir in Abschnitt 5 anhand zweier Beispiele im Bereich Konsumenten- und Investitionsentscheidungen mögliche Implikationen für das Controlling.

## 2. Erfahrungsbasierte und beschreibungsbasierte Entscheidungen

Entscheidungsforscher, gleichgültig ob in der Psychologie oder der Ökonomie, untersuchen die Frage, wie Menschen mit Risiko umgehen, häufig mit dem Werkzeug monetärer Lotterien. Die Lotterie, so die Annahme, ist das perfekte Double für reale Entscheidungssituationen. Genau wie Lotterien sind die realen Optionen, zwischen denen es zu entscheiden gilt, nichts anderes als Wahrscheinlichkeitsverteilungen über  $n$  mögliche monetäre (und/oder nicht-monetäre) Ausgänge. Daher unterstellen Entscheidungsforscher, wenn sie Personen dabei zusehen, wie sie sich zwischen Lotterien entscheiden, wie diese sich entscheiden würden, wenn es um die Wahl zwischen verschiedenen Investitionsfonds, Partnern oder Karrierewegen ginge. Man kann diese für die Entscheidungsforschung so zweckdienliche Annahme aus vielen Gründen kritisieren. Ein wichtiger Kritikpunkt ist allerdings dieser: Selbst wenn man die Rollenzuschreibung, wonach die Auswahl zwischen Lotterien das Sinnbild für Wahlentscheidungen schlechthin ist, akzeptieren würde, ist ein Aspekt besonders fragwürdig: Lotterien im Labor bestehen im Regelfall aus zwei oder mehreren Optionen und jede Option ist vollständig expliziert, das heißt, die möglichen monetären Ereignisse und dazugehörigen Wahrscheinlichkeiten werden vollständig beschrieben. Abb. 1 illustriert eine einfache und vollständig beschriebene Lotterie mit zwei Optionen: Eine der Optionen ist ein sicheres Ereignis (3 € mit Sicherheit), die andere Option offeriert eine relativ beträchtliche Summe von 32 € mit einer geringen Wahrscheinlichkeit von 10 % oder das Ereignis 0 € mit einer Wahrscheinlichkeit von 90 %.

Den Luxus, die Konsequenzen und Wahrscheinlichkeiten der Entscheidungsoptionen auf einem Silbertablett dargeboten zu bekommen, gibt es aber nur selten außerhalb des Labors. Das bedeutet jedoch nicht, dass reale Entscheidungen

notwendigerweise ohne eine Vorstellung von den zugrunde liegenden möglichen Ereignissen und Wahrscheinlichkeiten getroffen würden. Nicht selten stehen Erfahrungen – die eigenen oder die anderer Personen – in ähnlichen Situationen zur Verfügung. Und manchmal kann man auch die Entscheidungsoptionen erst explorieren, bevor die Entscheidung getroffen wird. Wein, den man beim Erzeuger kauft, kann man zum Beispiel vorher probieren. Oder man kann die Erfahrung anderer mit einem Hotel, einem Sportstudio oder einer Kinderkrippe zu Rate ziehen, bevor man sich für ein Angebot entscheidet. Entscheidungen, die auf Basis begrenzter und abzählbarer Erfahrungen mit einer Option getroffen werden, nennen wir **erfahrungsbasierte Entscheidungen (Decisions from experience)**. Im Gegensatz dazu nennen wir Entscheidungen, für die eine vollständige Beschreibung der möglichen Ereignisse und ihrer Wahrscheinlichkeiten vorliegt, **beschreibungsbasierte Entscheidungen (Decisions from description)**. Beschreibungen können symbolische oder graphische Form annehmen.

Die Analyse beschreibungs- und erfahrungsbasierter Entscheidungen existierte lange unabhängig voneinander. Erst in den letzten zehn Jahren hat sich dies grundlegend geändert. Die Initialzündung lieferte die Entdeckung eines systematischen und gravierenden Unterschieds in den Entscheidungen, die auf der Grundlage von Beschreibung und Erfahrung getroffen werden: die „Kluft“ zwischen beschreibungs- und erfahrungsbasierten Entscheidungen (Description-Experience Gap; vgl. Hertwig/ Erev, 2009).

## 3. Die Kluft zwischen beschreibungs- und erfahrungsbasierten Entscheidungen

Wie kann man das in Abb. 1 dargestellte Entscheidungsparadigma in ein erfahrungsbasiertes

### Wähle zwischen

A: 3 € mit Sicherheit

### oder

B: 32 € mit Wahrscheinlichkeit 10%

0 € mit Wahrscheinlichkeit 90%

Abb. 1: Eine typische beschreibungsbasierte Lotterie mit zwei Optionen



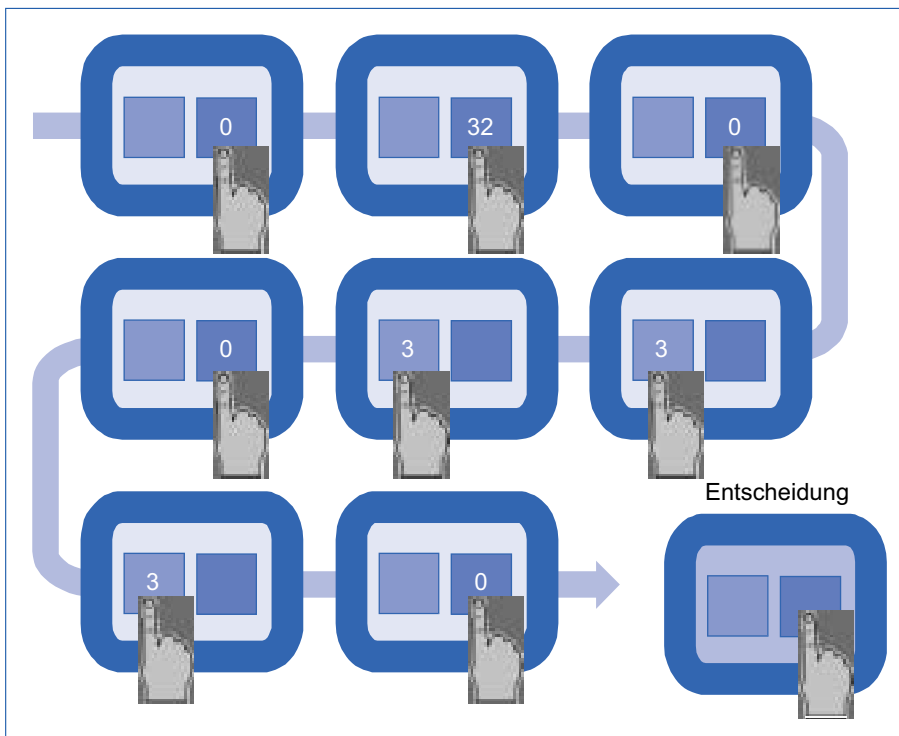


Abb. 2: Schematische Darstellung des Sampling-Paradigmas zur Untersuchung von entscheidungsbasierten Risikowahlen

rungsbasiertes Paradigma übersetzen, und zwar so, dass beide Situationen, zumindest im Prinzip, die gleichen Informationen zur Verfügung stellen? Entscheidungsforscher bedienen sich dazu der folgenden Anordnung: Versuchspersonen sehen zwei (oder mehr) Kästchen auf einem Computerbildschirm. Diese repräsentieren zwei zu Beginn völlig unbekannte Wahrscheinlichkeitsverteilungen. Das Klicken auf die Kästchen löst jeweils eine Zufallsziehung eines möglichen Ereignisses gemäß der zugrunde liegenden Verteilung aus. Basierend auf dieser Anordnung kann man drei Paradigmen unterscheiden. Im Sampling-Paradigma (Abb. 2; vgl. Hertwig et al., 2004) können Versuchspersonen so viele Zufallsziehungen auslösen, wie sie möchten, bevor sie die Exploration der Verteilungen beenden. Dann werden sie gebeten, sich zu entscheiden, auf welche Option sie setzen wollen. Diese letzte Wahl wird am Ende des Experiments ausgespielt und das Ergebnis bestimmt den monetären Gewinn oder Verlust der Person.

Das **Sampling-Paradigma** stellt eine Umwelt dar, in der die Suche nach Information möglich ist, ohne dass die zufällig gezogenen Ereignisse bereits zu materiellen Gewinnen oder Verlusten führen. Das

heißt, die Informationssuche zieht keine Kosten nach sich (mit Ausnahme von Opportunitätskosten). Daher könnte man die Verteilungen sorgfältig explorieren, bevor man eine Entscheidung trifft – so wie man sich beispielsweise sehr genau die Speisekarten von zwei teuren Gourmettempeln anschaut, bevor man sich für einen entscheidet. Im Sampling-Paradigma ist jener Zielkonflikt suspendiert, der unvermeidlich in Umwelten auftritt, in denen man sich zwischen unbekannten Optionen entscheiden muss: Soll man auf das Lernen neuer Informationen und langfristige Gewinnmaximierung setzen oder auf die Maximierung des Gewinns, basierend auf der Grundlage bekannter Informationen (**exploration-exploitation tradeoff**; vgl. Gupta et al., 2006). Das Sampling-Paradigma lässt sich jedoch schnell in ein Paradigma verändern, in dem dieser Zielkonflikt unvermeidbar ist. Das **Partial-Feedback-Paradigma** folgt der gleichen Logik wie das Sampling-Paradigma, mit dem einzigen Unterschied, dass nun jede Zufallsziehung (und deren Gesamtanzahl wird jetzt fixiert) bereits zum monetären Endergebnis beiträgt (d. h. dieses reduziert oder erhöht). Das dritte und letzte Paradigma ist das **Feedback-Paradigma**. Es ist so strukturiert wie das Partial-Feed-

back-Paradigma, nur dass jetzt nach jeder Zufallsziehung Informationen darüber gegeben werden, welcher Gewinn oder Verlust eingetreten wäre, wenn die andere Option gewählt worden wäre.

Bei der vergleichenden Analyse von beschreibungs- und erfahrungsbasierten Entscheidungen stehen diese drei Paradigmen und die folgende Frage im Fokus: Findet man unter Beschreibung (vgl. Abb. 1) und Erfahrung ein ähnliches oder systematisch unterschiedliches Entscheidungsverhalten? Abb. 3 illustriert beispielhaft die Antwort, die in einer Vielzahl von Studien gefunden wurde. Erfahrungs- und beschreibungsbasierte Formate führen nicht zu identischem Verhalten. Der Unterschied wird besonders deutlich, wenn eine Entscheidung zwischen einer riskanten Option, die entweder einen relativ hohen Gewinn oder hohen Verlust mit geringer Wahrscheinlichkeit (< 20 %; das seltene Ereignis) bietet, und einer sicheren Option, welche mit Sicherheit nur einen moderaten Gewinn oder Verlust garantiert, ansteht. Vor diese Wahl gestellt, fällt in allen drei erfahrungsbasierten Paradigmen die Wahl der Mehrheit auf die riskante Option, sofern das seltene Ereignis unattraktiv ist, aber auf die sichere Option, sofern das seltene Ereignis attraktiv ist (vgl. Abb. 3). Bei beschreibungsbasierten Entscheidungen findet sich die nahezu umgekehrte Mehrheitspräferenz. Insgesamt lässt sich der Unterschied so zusammenfassen: In erfahrungsbasierten Situationen entscheiden sich Menschen so, als ob der Einfluss der seltenen Ereignisse untergewichtet würde (relativ zu den objektiven Wahrscheinlichkeiten), während bei beschreibungsbasierten Entscheidungen ihr Einfluss übergewichtet zu sein scheint.

#### 4. Was verursacht die Kluft zwischen beschreibungs- und erfahrungsbasierten Entscheidungen?

Es gibt mehrere Faktoren, die verantwortlich für den Unterschied zwischen Erfahrung und Beschreibung sind. Die wichtigsten sind die frugale Informationssuche und die Art und Weise, wie die verfügbare Information verarbeitet wird. Beide Faktoren werden im Folgenden näher erläutert.

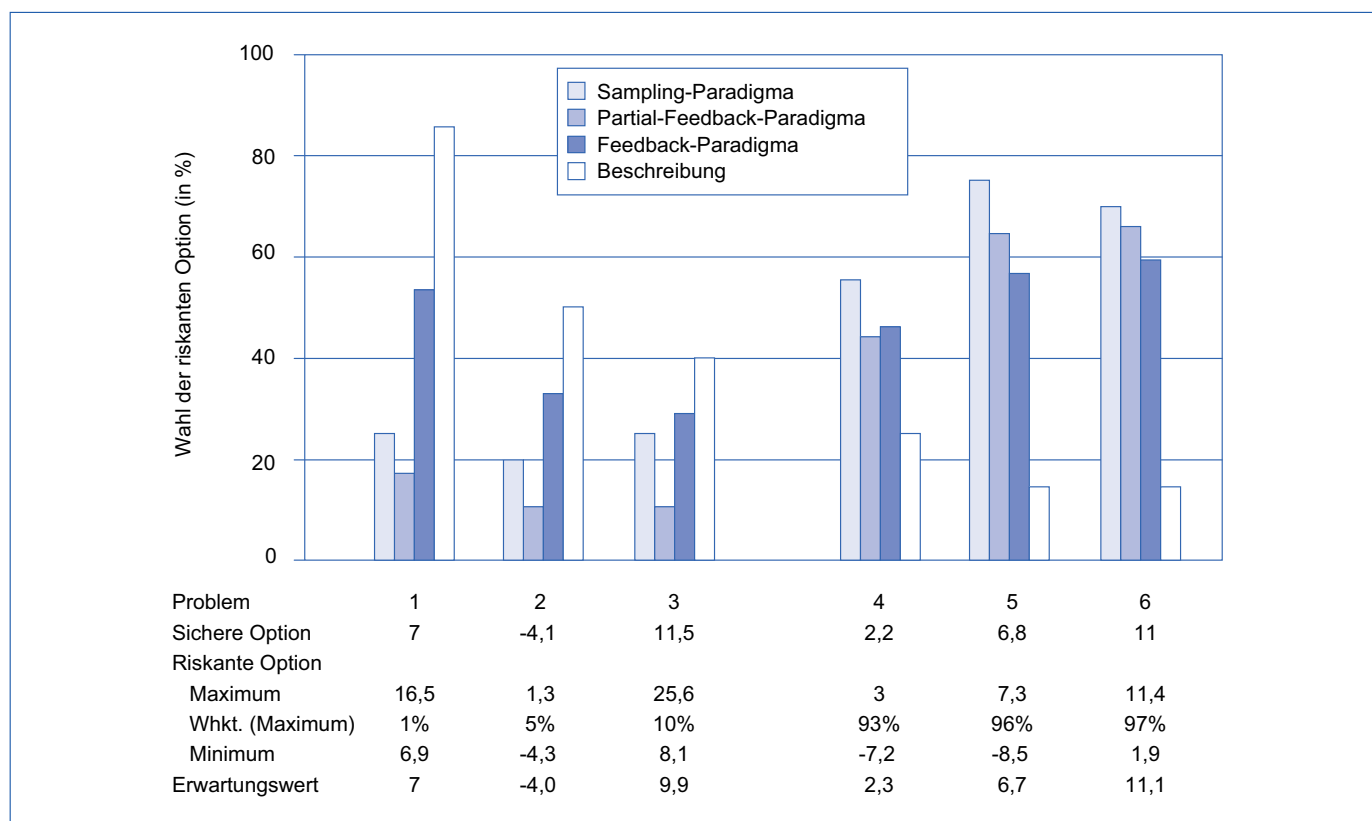


Abb. 3: Prozentsatz von Wahlen der mit Risiko behafteten Option in sechs verschiedenen Lotterien und als Funktion der vier experimentellen Paradigmen (entnommen aus Hertwig/Erev, 2009, S. 519)

## Frugale Suche

Das Sampling-Paradigma (vgl. Abb. 2) stellt es jeder Person frei, die Optionen sorgfältig zu explorieren (d. h. Zufallsstichproben zu ziehen). Im Durchschnitt findet man allerdings, dass die Anzahl der Ziehungen erstaunlich begrenzt ist. Abb. 4 zeigt die Ergebnisse einer Meta-

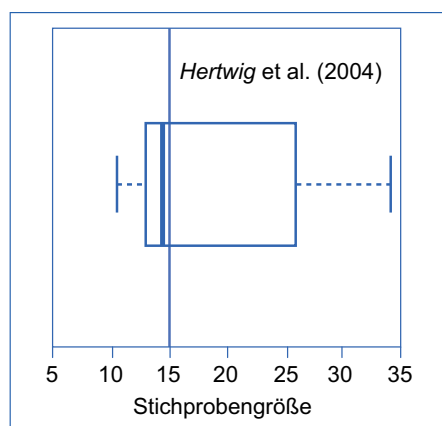


Abb. 4: Verteilung der Stichprobengrößen im Sampling-Paradigma (Ergebnisse aus 21 unabhängigen Datensätzen, basierend auf einer laufenden Metaanalyse von Dirk U. Wulff und Ralph Hertwig)

analyse von über 10.000 Entscheidungen, die von über 1.000 Versuchspersonen stammen. Der Median der Stichprobengröße in der Metaanalyse liegt sehr nahe an dem Median, der bereits in der ersten Untersuchung mit dem Sample-Paradigma beobachtet wurde (Median = 15; vgl. Hertwig et al., 2004). Im Durchschnitt zeigen sich Versuchspersonen mit relativ wenigen Ziehungen zufrieden und treffen bereits nach etwa  $7 \pm 2$  Beobachtungen pro Option eine Entscheidung.

Welche Konsequenzen hat diese relativ frugale Exploration? Stichprobenbasierte Schätzungen über die Wahrscheinlichkeit, mit der Ereignisse auftreten, sind mit einem Schätzfehler behaftet. Dieser Schätzfehler wird, ceteris paribus, desto grösser, je kleiner die Stichprobe ist. Im Extremfall könnte eine kleine Stichprobengröße sogar bedeuten, dass ein seltenes Ereignis darin überhaupt nicht zutage tritt und in der Entscheidung unberücksichtigt bleiben muss. Unterstellt man zum Beispiel ein seltenes Ereignis mit einer Wahrscheinlichkeit von 10 %, dann beträgt die „Gefahr“, dass das Ereignis in einer Stichprobe von sieben Ziehungen überhaupt nicht erscheint, beina-

he 50 %. Aber selbst wenn das Ereignis auftritt, kann eine kleine Stichprobe die Häufigkeit dieses Ereignisses trotzdem „unterrepräsentieren“. Das hängt mit einer Besonderheit des zugrunde liegenden stochastischen Prozesses zusammen. Die Verteilung der Häufigkeiten eines Ereignisses, beschrieben durch die Binomialverteilung, ist für kleine Wahrscheinlichkeiten (das Ereignis ist selten) und kleine Stichproben schief. Dies bedeutet, dass im Aggregat kleine Erfahrungsstichproben zu Entscheidungen führen können, in denen es mehr Leute gibt, die dem seltenen Ereignis zu wenig Gewicht beimessen, als Leute, die ihm zu viel Gewicht beimessen (im Hinblick auf seine objektive Wahrscheinlichkeit) – entweder weil es überhaupt nicht beobachtet wurde oder weil es durch das Bullauge einer kleinen Stichprobe noch seltener erscheint, als es objektiv ist.

Warum verlassen sich Menschen auf kleine Stichproben? Bevor dieses Verhalten zu schnell als „irrational“ abgehakt wird, sei Folgendes angemerkt. Die Analyse einer Computersimulation mit 1.000 (zufällig generierten) Lotterienproblemen zeigte, dass ein Agent mit lediglich sieben

Ziehungen aus jeder Option (14 insgesamt) im Schnitt bereits eine 81 %ige Chance hat, die Option mit dem höheren Erwartungswert zu erkennen. Danach nimmt der Zugewinn an Information durch weitere Ziehungen schnell ab (vgl. *Hertwig/Pleskac*, 2010). Die Nutzung frugaler Stichproben könnte also auch in der richtigen Intuition begründet sein, dass wenige Ziehungen bereits einen großen Informationsgehalt haben und der marginale Zugewinn weiterer Ziehungen schnell immer kleiner wird. Menschen können eine solche intuitive Kosten-Nutzen-Abwägung treffen. Darauf deutet die Beobachtung hin, dass eine deutliche Steigerung des monetären Anreizes, die attraktivere Option zu erkennen, auch zu deutlich mehr Exploration führt (vgl. *Hau et al.*, 2008). Dies ist aber bei Weitem nicht der einzige Faktor, der die Länge der Informationssuche und damit die Kluft zwischen erfahrungs- und beschreibungsbasierten Entscheidungen beeinflusst. Weitere Faktoren sind beispielsweise die Anwesenheit „wachsamer“ Emotionen (wie z. B. Furcht), das Aspirationsniveau einer Person sowie das Alter (und damit die abnehmenden kognitiven Kapazitäten; vgl. *Hertwig*, im Druck).

### Informationsverarbeitung

Neben der frugalen Suche gibt es einen zweiten Faktor, der gleichfalls zu einer augenscheinlichen „Untergewichtung“ seltener Ereignisse in erfahrungsbasierten Entscheidungen beiträgt. In beschreibungsbasierten Entscheidungen stehen alle Informationen gleichzeitig zur Verfügung (vgl. *Abb. 1*). Erfahrungsbasierte Informationen reihen sich entlang einer Zeitschiene auf (vgl. *Abb. 2*). Dieser Formatunterschied ermöglicht eine andere Integration der Information, womöglich

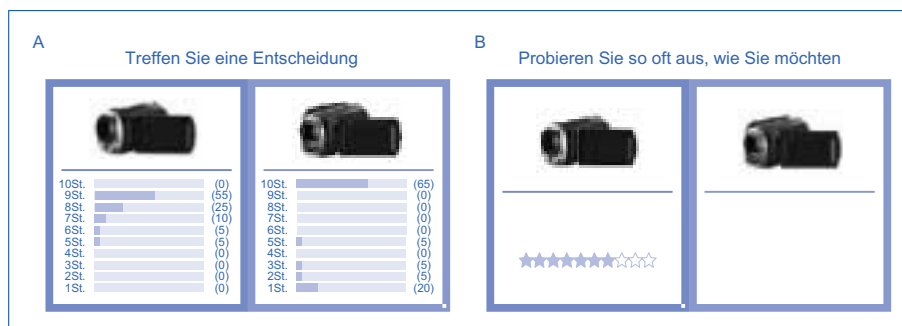
bedingt er sie sogar (vgl. *Hertwig*, im Druck). Zum Zeitpunkt der Entscheidung müssen beispielsweise die einzelnen Ziehungen oder eine komprimierte Form davon aus dem Gedächtnis abgerufen werden. Häufig materialisiert sich hier der sogenannte **Recency-Effekt**: Jüngst (recent) gesammelte Informationen erhalten ein höheres Gewicht in der Entscheidung als länger zurückliegende Informationen. Dies könnte zum einen mit durch die Funktionsweise unseres Gedächtnisses verursacht sein: Länger zurückliegende Ereignisse werden in der Regel schlechter erinnert als jüngere Ereignisse. Zum anderen könnte der Recency-Effekt eine adaptive Reaktion auf Umwelten sein, die häufig nicht-stationär und kompetitiv sind. Unabhängig von seiner Verursachung führt der Recency-Effekt zu einer „Unterrepräsentation“ seltener Ereignisse (siehe Binomialverteilung), da die ohnehin schon kleine Stichprobe (siehe oben) sich durch diesen Gedächtnisfilter Recency noch weiter verengt.

Die Liste der psychologischen Faktoren, die die Kluft zwischen beschreibungs- und erfahrungsbasierten Entscheidungen forcieren, ließe sich noch fortsetzen und auch auf das Partial-Feedback-Paradigma und Feedback-Paradigma ausdehnen (vgl. *de Palma et al.*, im Druck). Die Liste umfasst emotionale und motivationale Voraussetzungen des Entscheidenden, nicht zuletzt aber auch den Einfluss von Suchstrategien. Nicht nur die Menge der gesammelten Information, sondern auch die Suchstrategie (d. h. die Art und Weise, wie nach ihr gesucht wird) scheint unsere Entscheidungen zu determinieren (vgl. *Hills/Hertwig*, 2010).

## 5. Welche praktischen Implikationen hat die Kluft zwischen beschreibungs- und erfahrungsbasierten Entscheidungen?

Viele Entscheidungen erlauben nur erfahrungsbasierte Entscheidungen, weil die Handlungsalternativen sich nicht in Gestalt tabellierter Konsequenzen und assoziierter Wahrscheinlichkeiten präsentieren. In Zeiten von Big Data und Einkaufen im Internet stehen uns allerdings immer häufiger Beschreibungen in Form von aggregierten Erfahrungen anderer zur Verfügung. Routine ist dies beispielsweise bereits im Bereich des electronic commerce. Online-Plattformen wie zum Beispiel *Amazon* oder *Ebay* offerieren Bewertungstools. Nach dem Erwerb eines Produkts können Konsumenten quantitative Bewertungen beispielsweise in Form von Sternen abgeben. Diese Bewertungen wiederum helfen anderen potenziellen Kunden, die Qualität des Produkts (des Anbieters) einzuschätzen. Häufig spiegelt sich in den Bewertungen ein heterogenes Meinungsbild, was es nahelegt, diese Urteile in ein Gesamtbild zu integrieren. Dafür stehen dem Interessierten oft sowohl die individuellen Urteile als auch eine aggregierte Gesamtschau zur Verfügung. Auf *Amazon.com* zum Beispiel werden alle Bewertungen in Form eines Balkendiagramms komprimiert (ähnlich wie in *Abb. 5, A*). Dieses kann dann die Grundlage einer beschreibungsbasierten Produktentscheidung sein. Alternativ kann man sequentiell die Liste mit den einzelnen Bewertungen durchblättern und sich eine Meinung bilden (*Abb. 5, B*). Diese sequentiellen „Ziehungen“ können dann die Grundlage einer erfahrungsbasierten Entscheidung sein. Führen Balkendiagramm und die sequentielle Exposition mit individuellen Bewertungen zu identischen Entscheidungen? Die verfügbare Evidenz deutet darauf hin, dass Produktentscheidungen auf der Grundlage von Beschreibung bzw. sequentieller Erfahrung die gleichen systematischen Unterschiede aufweisen wie Wahlentscheidungen in Glücksspielen: Seltene Ereignisse (hier sehr positive oder negative Bewertungen) erhalten durch frugale Suche und Recency-Effekt zu wenig Gewicht in erfahrungsbasierten Entscheidungen (vgl. *Wulff et al.*, 2014).

Ist eines der beiden Informationsformate das offensichtlich bessere? Oder anders



**Abb. 5:** Beschreibungsbasierte (A) vs. erfahrungsbasierte (B) Konsumentenbewertungen (entnommen aus *Wulff et al.*, 2014)

gefragt: Sollte man nicht immer das beschreibungsbasierte Format präferieren, weil es eher vollständige und unverzerrte Information bietet als unsere (begrenzte) Erfahrung? Dass der Sachverhalt nicht so einfach liegt, belegt ein klassischer Befund der Forschung zu Anlegerpräferenzen: Relativ sichere Anlagen – wie zum Beispiel Rentenpapiere – werden trotz schlechterer, langfristiger Renditeerwartungen oft dem Erwerb von Aktien vorgezogen. Der Grund dafür scheint in der psychologischen Überbewertung des kurzfristig hohen Verlustpotentials der Aktien zu liegen (siehe Equity Premium Puzzle; vgl. *Bernatzi/Thaler*, 1995). Genau an jenem Punkt könnten erfahrungsbasierte Informationsformate von Vorteil sein. Eine aktuelle Studie fand, dass – im Vergleich zur graphischen Beschreibung des Risikos – ein Risiko-Tool, das auf der Präsentation sequentieller Ereignisse beruht und Anlegern die Möglichkeit gibt, Schritt für Schritt Erfahrungen mit der stochastischen Struktur von Aktienpreisen und Renditen zu simulieren, zu mehr Risikobereitschaft und gleichzeitig zu genaueren Einschätzungen des Erwartungswerts und der Wahrscheinlichkeit eines Verlustes führte (vgl. *Kaufmann et al.*, 2013). Eine größere Risikobereitschaft ist natürlich nicht a priori gut und die Anfangsbeispiele legen davon Zeugnis ab. Dennoch ist folgende Überlegung bedeutsam: „The use of experience sampling in financial simulations may be a fruitful strategy for banks to improve the quality of the information they provide about their investment products to ensure that clients understand both the risks they take and the amount of risk they are prepared to take“ (*Kaufmann et al.*, 2013, S. 336).

Ähnlich der Entscheidung für die richtige Investitionsanlage hängen viele unternehmerische Entscheidungen von der Kenntnis und Bewertung von Wahrscheinlichkeiten und dem Ausmaß der möglichen Konsequenzen einer Entscheidung ab. Damit drängen sich Implikationen der Kluft zwischen beschreibungs- und erfahrungsbasierten Entscheidungen auch für Risikomanagement in Organisationen auf. Viele Alltagsrisiken wie auch Risiken des unternehmerischen Handelns entziehen sich einer einfachen Risikoanalyse. Wahrscheinlichkeiten können oft nicht oder nicht genau quantifiziert werden, weil die in Frage stehenden Ereignisse nicht tabelliert wurden (die Risiken

waren nicht als solche erkennbar), singular sind oder die Stichprobe vergleichbarer Ereignisse sehr klein ist. In Fällen, in denen keine quantifizierten Größen vorliegen, scheint es naheliegend, dass Entscheidungsträger auch ihre persönlichen Erfahrungsstichproben zu Rate ziehen. In diesen Stichproben sind seltene Ereignisse – und dies liegt in der Natur der Sache – in der Regel unterrepräsentiert und werden in erfahrungsbasierten Urteilen vermutlich zu wenig Gewicht erfahren. Allerdings gilt auch dies: Ist ein seltenes Ereignis gerade kürzlich aufgetreten, wird dieses Risiko in erfahrungsbasierten Urteilen für einen gewissen Zeitraum vermutlich zu viel Gewicht erfahren.

Aber selbst wenn das Risikomanagement unternehmerische Chancen und Risiken identifiziert, tabelliert und quantifiziert, stellt sich die Frage, wie sich beschreibungs- und erfahrungsbasierte Beurteilungen zueinander verhalten. Schenken Entscheidungsträger ausschließlich beschreibungsbasierten Analysen von festgestellten Risiken Glauben oder werden diese auch durch die Filter der persönlichen Erfahrungen Neubewertet oder gar „verdrängt“? Erste Ergebnisse im Kontext von Warnhinweisen deuten darauf hin, dass bei seltenen Risiken die wiederholte Erfahrung, dass ein Risiko (bisher) nicht eingetreten ist, dazu führen kann, dass es möglicherweise nicht ernst genug genommen wird – ein Verhalten, das insbesondere auch im unternehmerischen Risikomanagement zu folgenschweren Fehleinschätzungen führen kann. Zum Beispiel findet man, dass Patienten ein Medikament mit einer schwerwiegenden, aber seltenen Nebenwirkung auch dann noch einnehmen, wenn vor den Nebenwirkungen gewarnt und das Medikament schlussendlich vom Markt genommen wurde. Die bislang guten Erfahrungen mit dem Medikament vermitteln ein falsches Gefühl der Sicherheit (vgl. *Barron et al.*, 2008). Zweifellos muss die zukünftige Forschung die interessante Interaktion zwischen Erfahrung und Beschreibung in der Bewertung von Risiken weiter ausloten und entschlüsseln.

Schlussendlich sei noch die folgende Überlegung erlaubt: Die Beschreibung der Wahrscheinlichkeit seltener Ereignisse führt eher zu deren Übergewichtung, wohingegen die Erfahrung der Wahrscheinlichkeit seltener Ereignisse eher deren Untergewichtung zur Folge hat. Da-

her stellt sich die Frage, ob gute Risikokommunikation – auch im Umfeld unternehmerischer Entscheidungen – darin bestehen könnte und sollte, den Entscheidungsträgern beide Formate zur Verfügung zu stellen. Neben den aggregierten, tabellierten Risiken könnte man „**experience sampling**“ als Methode verwenden, um – ähnlich wie bei Investigationsentscheidungen (vgl. *Kaufmann et al.*, 2013, S. 336) – Risiken verständlich zu kommunizieren und die Erfahrung eines Risikos unmittelbar zu simulieren.

## 6. Fazit

Bei einer Vielzahl unserer Entscheidungen spielen seltene, aber folgenschwere Ereignisse eine wichtige Rolle. Die bislang etablierte Sichtweise war, dass seltene Ereignisse in Risikowahlen zu viel Gewicht erfahren (gemessen an ihrer objektiven Wahrscheinlichkeit). Diese Sichtweise ist nicht falsch, aber ihr Gültigkeitsbereich ist augenscheinlich begrenzter als bislang vermutet. Der Schlüssel dazu liegt in einer in jüngster Zeit viel beachteten Beobachtung, dem **Description-Experience Gap** (vgl. *Hertwig/Erev*, 2009). Der Ausgangspunkt der Forschung zu dieser Kluft ist dieser: Symbolische Informationen über Risiken in Form einer Wahrscheinlichkeit oder einer Aussage begegnen uns allorts in unserem privaten und professionellen Umfeld. Warnhinweise auf Zigarettenpackungen kommunizieren symbolisch das Risiko des Rauchens, wenngleich in nicht-quantifizierter Form: „Raucher sterben früher“ oder „Rauchen verursacht tödlichen Lungenkrebs.“ Mediziner kommunizieren die Vorteile und die Risiken von Krebscreening-Verfahren in Form von deskriptiven statistischen Informationen. Die formalisierte Risikoberichterstattung in Organisationen tut das gleiche, wenn sie den Entscheidungsträgern Risikoberichte als Entscheidungsgrundlage zur Verfügung stellt. Neben der symbolischen Beschreibung von Risiken erfahren Menschen Risiken aber häufig auch durch den Filter ihrer persönlichen Erfahrung. Dem (Risiko-)Controlling kommt in diesem Kontext die Aufgabe zu, die Risikobewertung und damit Entscheidungen zu versachlichen.

Entscheidend ist Folgendes: Viele Untersuchungen zu der Frage, wie Menschen im Angesicht von Risiko und Unsicherheit Entscheidungen treffen, zeigen, dass



die Vermittlung von relativ unwahrscheinlichen Risiken mittels symbolischer Darstellungen dazu führen kann, dass diesen mehr Gewicht eingeräumt wird als ihnen in Anbetracht ihrer objektiven Wahrscheinlichkeit zusteht. Sobald aber für die Reaktion auf seltene Risiken nicht nur die symbolischen Informationen, sondern auch die eigene Erfahrung herangezogen wird, werden sie relativ angemessen eingeschätzt – zumindest dann, wenn die Erfahrungsstichprobe sehr groß ist. Bei Ereignissen, die so selten sind, dass sie selbst in einer großen Erfahrungsstichprobe nicht auftreten – zum Beispiel eine Weltwirtschaftskrise oder der äußerst seltene Ausbruch eines Vulkans –, neigt unsere begrenzte Erfahrung dazu, das Risiko zu gering zu gewichten. Beschreibung und Erfahrung eines Risikos sind also nicht einfach nur die zwei Seiten derselben Medaille.

#### Keywords

- Decisions from experience and description
- Decisions under risk and uncertainty
- Exploration
- Psychology of rare events
- Risk communication

#### Summary

This article gives an introduction into the description-experience gap – the observation of a systematic difference in risky choice based on symbolic descriptions versus sequential experience. The gap is particularly pronounced with rare events. Rare events appear to receive too much (description) and too little (experience) weight, respectively. Reasons are discussed and implications presented.

#### Literatur

- Barron, G./Leider, S./Stack, J., The effect of safe experience on a warnings' impact: Sex, drugs and rock-n-roll, in: *Organizational Behavior and Human Decision Processes*, 106. Jg. (2008), H. 2, S. 125–142.
- Benartzi, S./Thaler, R. H., Myopic loss aversion and the equity premium puzzle, in: *The Quarterly Journal of Economics*, 110. Jg. (1995), H. 1, S. 73–92.
- Gupta, A. K./Smith, K. G./Shalley, C. E., The Interplay Between Exploration and Exploitation, in: *Academy of Management Journal*, 49. Jg. (2006), H. 4, S. 693–706.
- Hau, R./Pleskac, T. J./Kiefer, J./Hertwig, R., The Description-Experience Gap in Risky Choice: The Role of Sample Size and Experienced Probabilities, in: *Journal of Behavioral Decision Making*, 21. Jg. (2008), H. 5, S. 493–518.
- Hertwig, R., Decisions from experience, in: Keren, G./Wu, G. (Hrsg.), *Blackwell handbook of decision making*, im Druck.
- Hertwig, R./Barron, G./Weber, E. U./Erev, I., Decisions from experience and the effect of rare events in risky choice, in: *Psychological Science*, 15. Jg. (2004), H. 8, S. 534–539.
- Hertwig, R./Erev, I., The description-experience gap in risky choice, in: *Trends in Cognitive Sciences*, 13. Jg. (2009), H. 12, S. 517–523.
- Hertwig, R./Pleskac, T. J., Decisions from experience: Why small samples?, in: *Cognition*, 115. Jg. (2010), H. 2, S. 225–237.
- Hills, T. T./Hertwig, R., Information Search in Decisions From Experience. Do Our Patterns of Sampling Foreshadow Our Decisions?, in: *Psychological Science*, 21. Jg. (2010), H. 12, S. 1787–1792.
- Kaufmann, C./Weber, M./Haisley, E., The role of experience sampling and graphical displays on one's investment risk appetite, in: *Management Science*, 59. Jg. (2013), H. 2, S. 323–340.

Lichtenstein, S./Slovic, P./Fischhoff, B./Layman, M./Combs, B., Judged frequency of lethal events, in: *Journal of Experimental Psychology: Human Learning & Memory*, 4. Jg. (1978), H. 6, S. 551–578.

de Palma, A./Abdellaoui, M./Attanasi, G./Ben-Akiva, M./Erev, I./Fehr-Duda, H./Fok, D./Fox, C. R./Hertwig, R./Picard, N./Wakker, P. P./Walker, J. L./Weber, M., Beware of black swans, in: *Marketing Letters*, im Druck.

Renn, O., Das Risikoparadox: Warum wir uns vor dem Falschen fürchten, Frankfurt a. M. 2014.

Taleb, N. N., *The Black Swan: The Impact of the Highly Improbable*, New York 2007.

Tversky, A./Kahneman, D., Advances in prospect theory: Cumulative representation of uncertainty, in: *Journal of Risk and Uncertainty*, 5. Jg. (1992), H. 4, S. 297–323.

Viscusi, W. K., *Smoke-filled Rooms: A Post-mortem on the Tobacco Deal*, University of Chicago Press 2002.

Wulff, D. U./Hills, T. T./Hertwig, R., Online Product Reviews and the Description-Experience Gap, Manuscript submitted for publication 2014.

Literaturtipps aus dem [Online-Archiv](#) der CONTROLLING:

- **Controlling-Schwerpunkt „Risikomanagement“, Ausgabe 1/2013, S. 1–63.**
- Julia Gans, Stefan Kreil und Thomas Schild, Effiziente Risikokommunikation für mehr Transparenz und eine risikobewusste Entscheidungsfindung – Das Risikomanagement der SAP, Ausgabe 4/5/2012, S. 237–240.
- Ludwig Sedlmeier, Entscheidungsunterstützung anhand multikriterieller Lösungsverfahren im Controlling, Ausgabe 10/2013, S. 545–547.

# Dirk U. Wulff

Max Planck Institute for Human  
Development  
Lentzeallee 94  
14195 Berlin  
Germany

Phone: +49 / (0)30 824 06 475  
Fax: +49 / (0)30 824 9939  
Email: [wulff@mpib-berlin.mpg.de](mailto:wulff@mpib-berlin.mpg.de)  
Email: [dirk.wulff@gmail.com](mailto:dirk.wulff@gmail.com)  
<http://www.mpib-berlin.mpg.de>

## RESEARCH INTERESTS

I am interested in the cognitive underpinnings of how people search in internal and external environments. To this end I study and model information search in experienced-based decision making and free recall from memory. My broader interests include: Cognitive modeling, risky choice, information search, optimal foraging, aspiration levels, formal models of memory, bounded rationality and heuristics.

## PUBLICATIONS

- Wulff, D. U., Hills, T., Hertwig, R. (in revision). The impact of short- and long-run frames on Search and Choice in Decisions from Experience.
- Wulff, D. U., Hills, T., Hertwig, R. (2014). The description-experience gap and online product reviews.
- Van den Bos, W., Jenny, M., & Wulff, D. U. (2014). Open Minded Psychology.
- Hertwig, R., & Wulff, D. U. (2014). Risikoentscheidungen: Die Kluft zwischen Erfahrung und Beschreibung. *Controlling: Zeitschrift für Erfolgsorientierte Unternehmenssteuerung*.
- Wulff, D. U. Hills, T., Hertwig, R. (2013) Wormholes in Memory: Is memory one representation or many?. *Proceedings of the Cognitive Science Society*, 35.
- Wulff, D. U. Hills, T., Hertwig, R. (2012) Adaptive Information Search and Decision Making over Single and Repeated Plays. *Proceedings of the Cognitive Science Society*, 34.
- Wulff, D. U., & Meiser, M. (2010). *Stochastic Dependence in Source Memory: Testing Measurement Models of Retrieval Experiences (Unpublished diploma thesis)*. Philipps-University Marburg, Germany.

## MANUSCRIPTS

- Wulff, D., U., & Pachur, T., (in prep.). Interpreting Noise as Forgetting in Modeling Valuation From Experience? A Comment on Ashby and Rakow (2014)

Wulff, D. U., Mergenthaler Canseco, M., Hertwig R., (In prep.). Recency exists in the sampling paradigm of decisions from experience, but why?.

Schulte-Mecklenbeck\*, M., Wulff\*, D. U., Haslbeck, J. (In prep.). Is search like choice in decisions from experience. A process tracing study. \*shared authorship, order determined by coin flip

Wulff, D. U., Braun, M., Hills, T. T., & Khader, P. (in prep.). Neural correlates of dynamic search in memory.

Wulff, D. U., Hertwig, R., Mergenthaler-Canseco, M. (in prep.). The description-experience gap in the sampling paradigm: A meta-analytic review.

Wulff, D. U. Mata, R., Hills, T. T. (In prep.). Free memory search across domain and lifespan.

## EDUCATION & POSITIONS

2012-2014	Doctoral student at the Center for Adaptive Rationality (ARC), Max Planck Institute for Human Development, Berlin.
2010-2012	Doctoral student at the department for Cognitive and Decision Sciences, University of Basel
2004-2010	Diploma in Psychology at Philipps-University Marburg Focus on Psychological Methods, Cognitive Psychology and Clinical Psychology.

## AWARDS & GRANTS

2014	Acquired funding for the Summer Institute on Bounded Rationality, 198,000€
2013	Travel Grant of the Swiss National Science Foundation, \$3,200.
2013	Funding for the 6 <sup>th</sup> JDM Workshop for Young Researchers from the European Association for Decision Making (EADM). \$4000
2012	Travel Grant of the Swiss National Science Foundation, \$1,900.
2012	Travel Grant of the Swiss National Science Foundation, \$2,700.
2012	Invitation to the Summer School on Computational Modeling of Cognition \$1000
2012	Travel Grant of the Swiss National Science Foundation, \$2,700.
2011	Best Student Poster Award, 3 <sup>rd</sup> place, Annual Conference for Judgment and Decision Making, Seattle, 2011, \$250.
2011	Travel Grant of the Swiss National Science Foundation, \$2,000.
2011	Travel Grant of the Swiss National Science Foundation, \$700.
2011	Scholarship of the European Campus of Excellence, \$800.
2011	Invitation to the Summer School on Bounded Rationality of the Max Planck Institute for Human Development, \$500.
2010	Travel grant of the German Academic Exchange Service, \$500.

## PRESENTATIONS

- Wulff, D.U., Hills, T.T., & Hertwig, R. (2013). The Description-experience Gap in the sampling paradigm: a meta-analytic review. Paper presented at the Annual Meeting of the European Association of Social Psychology, Amsterdam, Netherlands.
- Wulff, D.U., Hills, T.T., & Hertwig, R. (2013). The Description-experience Gap in the sampling paradigm: a meta-analytic review. Paper presented at the 7<sup>th</sup> JDM workshop for Young Researchers at the University of Mannheim, Mannheim, Germany.
- Wulff, D.U., Hills, T.T., & Hertwig, R. (2014). Is memory one representation or many? Evidence for the dynamic use of multiple representations. Paper presented at the 56. Tagung experimentell arbeitender Psychologen (TEAP). Gießen, Germany.
- Wulff, D.U., Hills, T.T., & Hertwig, R. (2013). Long Versus Short-term Aspirations in Decisions from Experience. Paper presented at the 24th Subjective Probability, Utility, and Decision Conference (SPUDM). Barcelona, Spain.
- Wulff, D.U., Hills, T.T., & Hertwig, R. (2013). Wormholes in Memory: Is memory one representation or many?. Poster presented at Bayesian Modeling for Cognitive Science. A WinBUGS Workshop. Amsterdam, Netherlands.
- Wulff, D.U., Hills, T.T., & Hertwig, R. (2013). Wormholes in Memory: Is memory one representation or many?. Poster presented at the Annual Meeting of the Cognitive Science Society. Berlin, Germany.
- Wulff, D.U., Schulte-Mecklenbeck, M., & Haslbeck, J. (2013). Is search like search in decision from experience. A process tracing study. Paper presented at the 55. Paper presented at the 6<sup>th</sup> JDM workshop for Young Researchers at the Max Planck Institute for Human Development, Berlin, Germany.
- Wulff, D.U., Hills, T.T., & Hertwig, R. (2013). Online Product Reviews and the Description-Experience-Gap. Paper presented at the 55. Tagung experimentell arbeitender Psychologen (TEAP). Wien, Austria.
- Wulff, D.U., Hills, T.T., & Hertwig, R. (2012). A review and reanalysis of the sampling paradigm: The impact of problem characteristics on search. Poster presented at the Annual Meeting of the Society for Judgment and Decision Making. Minneapolis, MN, U.S.
- Wulff, D.U., Hills, T.T., & Hertwig, R. (2012). Adaptive Information Search and Decision Making. Paper presented at the Annual Meeting of the Cognitive Science Society. Sapporo, Japan.
- Wulff, D.U., Hills, T.T., & Hertwig, R. (2012). Search in Memory: How we use multiple cues. Oral presentation at the Summer School on Computational Modeling in Bergün, Switzerland.
- Wulff, D.U., Hills, T.T., & Hertwig, R. (2012). Time frame dependent information search and decision making. Paper presented at the 54. Tagung experimentell arbeitender Psychologen (TEAP). Mannheim, Germany.
- Wulff, D. U., Hills, T.T. (2012). Information Search in Decisions from Experience. Invited talk at DR@W Forum at University of Warwick, UK.



- Wulff, D. U., Hills, T.T., & Hertwig, R. (2011). Information Search in the Short and Long run. Poster presented at the Annual Meeting of the Society for Judgment and Decision Making. Seattle, WS, U.S.
- Wulff, D. U., Hills, T.T., & Hertwig, R. (2011). The Role of Decisions from Description and Decisions from Experience in Online Consumer Choice. Paper presented at the 23<sup>rd</sup> Subjective Probability, Utility, and Decision Conference (SPUDM). Kingston upon Thames. UK
- Wulff, D. U., Hills, T.T., & Hertwig, R. (2011). The Description-Experience Gap generalizes to Online Consumer Choice. Paper presented at the 4<sup>th</sup> JDM workshop for Young Researchers at the Max Planck Institute for Research on Collective Goods, Bonn, Germany.
- Wulff, D. U., Hills, T.T., & Hertwig, R. (2011). The Description-Experience Gap in Online Consumer Choice. Poster presented at the Summer Institute for Bounded Rationality at the Max Planck Institute for Human Development, Berlin, Germany.
- Wulff, D. U., Hills, T.T., & Hertwig, R. (2011). The Role of Decisions from Description and Decisions from Experience in Online Consumer Choice. Paper presented at the research colloquium for Decision and Economic Psychology. University of Basel, Switzerland
- Wulff, D. U., Meiser, T. (2011). Examining stochastic dependence in source memory for recognition memory processes. Poster presented at the 53. Tagung experimentell arbeitender Psychologen (TEAP). Halle, Germany.

## TEACHING

- |            |  |
|------------|--|
| 2011- 2012 | Undergraduate research seminars at University of Basel.  |
| July 2010  | Introductory R-Workshop for the psychological methods department at Philipps-University Marburg. |
| 2009-2011  | TA for statistics at Philipps-University Marburg.  |

## ACADEMIC VISITS

- |             |   |
|-------------|---|
| 2010        |   |
| Jan. – Mar. | Human Memory Lab at UC Davis, California. Host: Prof. Andrew Yonelinas.                         |
| 2009        |   |
| Mar.-May    | Department of psychological methods at the Philipps-University Marburg. Host Dr. Oliver Christ. |

## AD-HOC REVIEWER

- Journal of Behavioral Decision Making
- Journal of Experimental Psychology: Learning, Memory & Cognition

## WORKSHOPS ATTENDED

Bayesian Modeling for the Cognitive Science. A WinBUGS Workshop. University of Amsterdam, August 2013

SNF Summer School on Computational Modeling of Cognition, Bergün, Switzerland, June, 2012.

Summer Institute on Bounded Rationality at the Max Planck Institute for Human Development, Berlin. June, 2011.

Fate of the Memory Trace. Summer School of the European Campus of Excellence at the Ruhr-University Bochum. September, 2011.

Bayesian Modeling in the Cognitive Sciences. Workshop at the University of Zürich. December 2010.

Cognitive Modeling in R. Workshop at the Philipps-University Marburg. December, 2010.

## Supervision

Florian Brühlmann, University of Basel, Intern – “Searching memory for countries”

Max Mergenthaler Canseco, Free University Berlin, Intern – “The description-experience gap”

Moria Braun, University of Marburg, Master thesis – “EEG correlates of verbal fluency for countries”